The Dissertation Committee for Meaghan Suzanne Edmonds certifies that this is the approved version of the following dissertation:

Utilizing implementation data to explain outcomes within a theory-driven evaluation model

Committee:

_____

Gary Borich, Co-Supervisor

_____

Keenan Pituch, Co-Supervisor

_____

Edmund Emmer

_____

Gregory Roberts

_____

Sharon Vaughn

Utilizing implementation data to explain outcomes within a theory-driven evaluation model

by

Meaghan Suzanne Edmonds, B.A.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2007

Acknowledgments

As with any major accomplishment in life, the support and guidance of many were instrumental in obtaining this degree. I am blessed to be surrounded by such giving people. My heartfelt gratitude and appreciation go to…

My committee members, who each played an important role in my academic development and the formation of this dissertation. A special thanks to Dr. Vaughn, who has been a mentor and colleague throughout my graduate career.

My parents and family, who encouraged my pursuit of this degree and fostered my love of learning.

My children, Thomas and Baby Girl (due soon!), who serve as my inspiration and reminders of what is truly important in life.

My dear husband, Shannon, whose patience, love and belief in me make all things possible.

Utilizing implementation data to explain outcomes within a theory-driven evaluation model

Meaghan Suzanne Edmonds, Ph.D.

The University of Texas at Austin, 2007

Supervisors: Gary Borich, Keenan Pituch

This study examined the moderating effects of teachers' implementation of a research-based comprehension intervention on a related student outcome. In addition to looking at the utility of including implementation data in a model of student outcomes, the stability of implementation ratings across occasions and the relationship between two implementation data sources (teacher logs and researcher ratings) were examined. The program featured in the study consisted of research-based comprehension strategy instruction implemented in 4th grade classrooms during social studies. Two measures of implementation—fidelity and overall instructional quality—did not predict student outcomes. In the tested model, a student's comprehension skills upon entering 4th grade did more to predict post-intervention comprehension achievement than did the teacher's instructional practices. Secondary analyses showed that an overall measure of teacher quality appears to be relatively reliable across only a few measurement occasions. Fidelity scores were less stable across occasions. The alternative method of collecting implementation data used in this study (audio recordings) appears to offer a viable and less costly means of obtaining implementation data. In addition, when measured at a macro level, implementation fidelity data from two sources (teacher logs and researcher ratings) were moderately correlated. Results inform future theory-driven evaluation

activities by providing information on approaching the task of documenting implementation and using that information to understand program outcomes.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

An important component of program evaluation is monitoring program processes and implementation (Rossi, Freeman, & Lipsey, 1999). Theory-driven evaluation (Chen, 1990; Weiss, 1997a, Weiss, 1997) calls for documentation of program implementation as a means of better understanding how program enactment contributes to outcomes. Weiss (1997a) argues that because program evaluation has the potential to influence public policy, understanding the relationship between outcomes and practices believed to influence program outcomes (i.e., implementation, teacher quality) is a necessary goal of an effective evaluation.

In addition to determining program effects, policymakers and program stakeholders are often interested in determining the extent to which participants were faithfully implementing the components theorized to influence outcomes. In the context of instructional programs and interventions, syntheses of research consistently show that effect sizes are lower for teacher-implemented than researcher-implemented interventions (e.g., Edmonds et al., in press; Rosenshine & Meister, 1994; Swanson, 1999). This is likely because researchers adhere more closely to the intervention model as intended, whereas teachers may adapt the intervention to suit their teaching style, curricular constraints or student needs. This phenomenon makes it all the more important to examine the effects of varying or less than optimum implementation on program outcomes.

Both the What Works Clearinghouse (WWC, 2005) and the National Research Council's document On Evaluating Curricular Effectiveness (NRC OECE, 2004) include program implementation in their criteria for judging intervention effectiveness. Although

program theory and implementation are conceptualized differently within each document, both recognize the importance of accounting for implementation variation in interpreting study results.

Implementation monitoring can involve assessing not only the presence of essential program elements, but also the adequacy of those components. As is the case in the theory-driven evaluation model, monitoring often goes beyond simply determining compliance and examines the extent to which program components are implemented. Such analyses serve as a means of evaluating program outcomes within the context of real-world implementation. In classroom interventions, instructional practices are one component of implementation: other components include professional development and materials used. However, because the instruction is the most proximal program component to the first order outcome (student reading achievement), assessing fidelity to the "treatment" or instructional model as defined by the program would be a critical function of any related theory-driven program evaluation.

Although researchers are more likely to report information on implementation now than in the past, it is still rare that such data are examined for their moderating effects on outcomes. Those studies that do are usually retrospective studies of the existing curriculum as opposed to studies of a specified program or intervention. For example, reading research has examined implementation of the enacted curriculum through classroom observations in early elementary grades (e.g., Foorman & Schatsneider, 2003). This line of research has also examined the impact of overall teacher quality on reading outcomes (Foorman, Schatschneider, Eakin, Fletcher, Moats, & Francis, 2006; Taylor, Pearson, Clark & Walpole, 2000).

However, there is little information on the utility of documenting and analyzing instructional fidelity in programs where the instructional intervention is based on a relatively prescribed model. In a review of research on comprehension strategy instruction, Lysynchuk and colleagues (Lysynchuk, Pressley, Ailly, Smith & Cake, 1989) found that only 30% of reviewed studies reported implementation data, which they defined as time spent on the independent variable tasks. Moreover, there is little information on the moderating function of implementation and overall teacher quality in instructionally complex domains such as reading comprehension interventions.

One methodology for determining the level of implementation of a given program is through classroom observations. Although reading research often relies on observation methods to determine intervention implementation, the field lacks convincing and converging evidence on methods of determining implementation fidelity in cost-effective ways. Of particular interest is documenting implementation through less costly methods than direct observation. Conducting in-classroom observations for large scale research and evaluation projects can be cost prohibitive, necessitating the use of alternative methodologies for capturing information on instructional practices. For example, an ongoing large-scale evaluation of comprehension interventions funded by U.S.DOE (Title 1) has had to limit in-classroom observations to two in order to meet cost projections (Mathematica, 2006). The efficacy of alternatives such as teacher logs and surveys have been documented in the research to some extent, but less is known about more novel approaches such as audio-taping teachers' instruction.

This study examined the premise of documenting implementation as a necessary component of program evaluation by exploring the moderating effects of teachers'

implementation of a research-based comprehension intervention on related outcomes. In addition to looking at the utility of including implementation data in a model of student outcomes, alternative methods of obtaining implementation data were used. The program featured consisted of research-based comprehension strategy instruction implemented in 4th grade classrooms during social studies.

A major goal of the study was to determine whether implementation and teacher quality are related to student outcomes within a comprehension intervention program. A secondary goal was to examine different means of collecting this type of data. The variance in student outcomes explained by two types of implementation data—fidelity and overall teacher quality—and the correlation between teacher-provided and researcher coded implementation data was examined to inform future program evaluation endeavors that include implementation as a component of a theory-driven model. Research questions were:

1) Do classrooms vary in comprehension achievement?
2) To what extent do ratings of implementation fidelity and overall teacher effectiveness affect student comprehension outcomes after controlling for initial reading status?
3) Is there an interaction between implementation fidelity and teacher effectiveness on student comprehension outcomes after controlling for initial reading status?
4) What is the correlation between teacher logs and research-rated protocols on length of session (dosage estimate) and number of components implemented for sample lessons?

5) Are estimates of teachers' implementation stable across occasions?
6) What innovations do teachers make to the program and what are the characteristics of teachers with high class average posttests?

Although the study examined a reading comprehension program, the intent was not to evaluate the program itself, but rather to examine the role of implementation in explaining differences in outcomes across teachers. Results inform future theory-driven evaluation activities by providing information on how to approach the task of documenting implementation and use that information to understand program outcomes.

Chapter 2: Literature Review[1]

Theory-driven evaluation

In the broadest sense, program theory is a model of how the program or instructional intervention is intended to work (Weiss, 1997; Yin, 1994). Theory-driven evaluation is not an evaluation of whether an intervention works, but *how* the program works in the real world context (Rogers, 2000). Most often the application of theory-driven evaluation involves a model of program activities and mechanisms (those behaviors or practices theorized to influence results) and an evaluation of outcomes using that model. Although such models do account for real-world implementation, the models themselves are usually perceived to generalize to implementation in any context. In this study, the evaluation of outcomes is related to a particular theory of comprehension—a theory of how students understand what they read and the instructional mechanisms believed to bring about improved comprehension.

Three primary components of a program theory model include the intervention or program as intended, measures of the implemented intervention and resultant achievement outcomes (Rogers, 2000). Program theory models can be quite intricate, increasing in complexity as the number of program components believed to affect change also increases. However, many interventions lend themselves to relatively simple theoretical models that consist of program activities and the primary mechanism believed

---

[1] Parts of the literature review were adapted from: Edmonds, M.S. (2005). *An Examination of the Properties of a Classroom Observation Instrument Purported to Measure the Proportion of Time Allocated to Research-based Reading Instruction*. A Master's report submitted to the Faculty of the Graduate School of The University of Texas at Austin.

to bring about the desired change in student achievement. A simple program model describes the intended (the lesson components), the enacted (implementation) and the achieved (student outcomes) (Rogers, 2000; Weiss, 1990).

Conceptualizing implementation differs across evaluation entities. For example, the What Works Clearinghouse (WWC, 2005) infers high implementation fidelity in the absence of strong evidence to suggest the contrary. In this case implementation is operationalized as a thorough intervention description that would allow for replication and the assumption of implementation fidelity if there is no evidence to suggest that groups within the experimental condition had different experiences. On the other hand, the On Evaluating Curricular Effectiveness framework (OECE) used by the NRC emphasizes a stricter definition of program theory by calling for documentation of elements such as professional development, teacher effects, and resources, among others. In fact, the OECE guidelines recommend that at a minimum, intervention reports include critical elements of a program theory model, such as professional development contact hours and a measure of the extent to which the materials were used, before an intervention can be judged on efficacy (Confrey, 2006).

When interventions are not standardized or highly prescribed or there is reason to believe that fidelity to a specified intervention is varied, then consideration of implementation data is necessary to better understand outcomes. Without such data, evaluators cannot claim with confidence that the outcomes resulted from the program as intended. A failure to document and account for implementation fidelity threatens the internal validity of any program evaluation.

By looking at implementation as one part of a program theory, the researcher recognizes that the context of classrooms and teachers' individual decisions, choices and preferences about pedagogy and instructional practices will likely result in an enacted curriculum that differs from the intended. These differences also vary from slight modifications to considerable deviations. Based on studies reviewed, the NRC concluded that there evidence to suggest that "variation within an implementation of a curriculum [is] substantial" (pg. 206, Confrey, 2006). One can argue that without information on treatment fidelity, researchers fail to adequately attend to important variables that may moderate program effects. In addition, by inferring high implementation without documenting or examining implementation data, an opportunity to better understand the everyday practices of program implementers is lost.

Comprehension theory and instruction

The study presented here was part of a larger Institute of Education Sciences-funded[2] study being conducted by The Vaughn Gross Center for Reading and Language Arts, in partnership with Texas A&M University. The larger study was designed to address the critical need to better understand the "4th grade slump" by focusing on comprehension of content-area text. The "4th grade slump" refers to the phenomenon of declining reading achievement once students enter the upper elementary grades. Data trends indicate that some students who read proficiently in the lower elementary grades experience comprehension difficulties upon entering 4th grade, likely

[2] Research referred to hereafter is funded by the U.S. Department of Education's Institute of Educational Sciences, grant contract number R305M050121A (*Enhancing the quality of expository text instruction and comprehension through content and case-situated professional development*)

as a result of encountering increasingly complex texts with more sophisticated vocabulary (Chall, 1983).

The larger study was designed to address questions related to the effect of teacher-implemented, evidence-based practices on 4th-grade students' comprehension and vocabulary. The research project is intended to result in evidence on how professional development on research-based practices works in context to improve the quality of teachers' instruction and students' reading comprehension. The study presented here, a sub-analysis of data, focuses only on implementation data from the participating comprehension intervention classrooms.

A theoretical model of comprehension. To examine the effects of implementation, it is necessary to first explicate the theory of comprehension that informs the instructional program and the evaluation model. Comprehension, one of several components of the reading process, is the ultimate goal of reading. Comprehension refers to the ability to understand information presented in text (Pressley, 2002). Despite the succinct definition of comprehension, it is a complex skill with multiple levels. Kintsch's (1998) construction-integration (CI) theory of comprehension differentiates between three comprehension levels. At the most basic level, comprehension consists of decoding processes. Students successful at this level are able to comprehend at the word level, producing what Kintsch refers to as "meaning units" (Kintsch & Kintsch, 2005). These units consist of words, phrases and sentences. At the next level, readers move to text analysis, garnering information and building a text-based model. In the final level, a reader *learns* from text by actively constructing meaning during reading and connecting new information with prior knowledge (Deschler & Hock, 2007; Kintsch & Kintsch,

2005). It is at this highest level of comprehension that readers build what Kintsch refers to as a situation model, a model of the information that places new text information within the schema of existing knowledge and ideas. Mirroring cognitive psychology's theories of integration and cognitive dissonance, the CI theoretical model progresses from context-free, text-based meaning construction towards the integration of new information, at which point a new mental structure is developed and learning has occurred.

To address these various levels of understanding, comprehension *instruction* must be multifaceted and systematic. Effective comprehension instruction is more than simply asking questions to gauge students' understanding of text; it involves explicitly modeling and teaching comprehension strategies so that children are able to independently learn from text. This type of comprehension instruction is usually characterized by teaching students specific strategies to help them better grapple with the meaning of text.

Swanson (1999) showed a large effect (d = .72) for comprehension interventions in general for students with reading disabilities. The effects were even higher when the intervention included derivatives of direct instruction. Further, strategy cuing (i.e., interventions that teach and then cue students to apply comprehension strategies) explained significant variance beyond characteristics of what the meta-analysts defined as the "core comprehension model" (Swanson, 1999). Dole, Brown, and Trathen (1996) also found superior gains in comprehension for students who received strategy instruction over those who received content or traditional comprehension instruction. To address the levels of comprehension represented by the CI theoretical model, instruction is beneficial when teachers scaffold students' comprehension skills and teach a range of strategies

from those that develop text-based models to that that help students create the more complex situation models indicative of deeper understanding.

Research-based comprehension strategy instruction that addresses the levels in the CI model includes previewing the text, preteaching key ideas and words, and showing students how to answer and generate questions, monitor their understanding of text, summarize what they have read, and use graphic organizers (Duke & Pearson, 2002; NRP, 2000; Pressley, 2002). The intervention featured in this study focuses on reading strategy instruction during social studies. Therefore, the following review addresses strategies related to expository text comprehension and content area instruction.

Previewing. Some research has provided evidence that activating relevant knowledge prior to reading content-area texts may enhance memory and understanding of what was read (NRP, 2000). Typical previewing activities activate students' prior knowledge by asking students to report what they already know about the subject and what they will likely learn based on clues provided by text features (e.g., pictures, subheading, titles). However, students with little prior knowledge of a subject likely require more teacher-directed previews that provide relevant background information and highlight key ideas. Such practices have been shown to be effective. For example, Graves and colleagues (1983) reported that teacher-directed previews that provide students with information such as key ideas and definitions of unfamiliar words can improve information recall. The previews featured in this study differed from typical previewing activities that elicit information from students without explicitly providing adequate background information (Graves, Cooke, & Laberge, 1983).

For upper elementary students, a particular stumbling block at the most basic comprehension level is decoding complex, content-specific words. Content-area texts are rife with multi-syllable proper nouns that may be unfamiliar to students. Pre-teaching proper nouns (both what they mean and how to say them) can remove these potential impediments to comprehension (Fletcher et al., 2006).

Questioning. Ample evidence supports the efficacy of providing support for questioning strategies through opportunities for students to ask and answer their own questions about the text (NRP, 2000). There is strong evidence to support explicitly teaching students questioning strategies as part of an instructional program, including content area instruction. Students who are taught to ask themselves and others questions about what they read demonstrate improved reading comprehension (Rosenshine, Meister & Chapman, 1996; Wong & Jones, 1982). The efficacy of teaching students to generate their own questions has been more consistently reported in the literature than has been question answering strategies. In two syntheses that reviewed question-generating strategies, high effects were found. For example, Mastropieri and colleagues (1996) found an overall effect of $d = 1.33$ for questioning strategies used with struggling readers (Mastropieri et al., 1996). In a well-known synthesis of questioning interventions, Rosenshine, Meister and Chapman (1996) found consistently large effects (d =.85-.95).

Effective questioning strategy instruction teaches students to generate their own questions, as opposed to the teacher primarily asking and answering all questions. Teaching students self-questioning strategies can extend comprehension to high levels of understanding and help both teachers and students monitor how well the text is understood. The types of questions asked and generated play an important role in strategy

instruction as some questions, such as lower level questions, can limit students thinking (Raphael et al., 2006).

Raphael (1986) developed and studied Question-Answer Relationships (QARs), a model to teach students how to answer different question types. Question types range from "Right There" in the text, which require only a very basic understanding of text, to "The Author and You," which require students to think about what they have read and make connections to their prior knowledge and experiences.

Interactive questioning, which involves guided discussions with teacher feedback on student-generated questions and answers, has been widely studied (Beck, McKeown, Hamilton, & Kucan, 1997; Beck, McKeown, Worthy, Sandora, & Kucan, 1996). This research has shown that students who participate in Questioning the Author, which involves engaging students with text through discussions and feedback, tend to view reading as a means of learning new information rather then a task to be completed (Beck & McKeown, 2001; Pressley, 2001). However, guiding a high-quality discussion that leads to student engagement with the ideas in text can be challenging for even the most highly skilled teachers (Beck & McKeown, 2001).

Main idea and summarization. Effective comprehension instruction also includes teaching students to write important ideas about what they've read and to summarize these ideas across passages or paragraphs. While summarization has sufficient evidence of efficacy in improving comprehension, it is not clear whether there are differential effects of this practice across text types. Main idea instruction is most often subsumed under summarization instruction.

Summarization strategies include first teaching students to write important ideas about what they've read (i.e., main ideas). Later, after students gain proficiency in writing main idea statements, these statements are combined into high quality summaries. Students are often asked to summarize text without being given a model or guidelines on how to best approach the task. Guidelines for effective summarization include (NRP, 2000) (a) identify or create a topic sentence using the "big idea" from the text (b) combine main idea statements, (c) delete redundant information, and (d) use succinct language (as few words as possible) when generating statements and summaries.

As mentioned, students are often taught how to identify the main idea for a smaller piece of text before learning how to create a summary. Research has shown that systematic and explicit instruction in how to identify a main idea can lead to improved comprehension (Graves, 1986; Jenkins, Heliotis, Stein, & Haynes., 1987; Jitendra, Cole, Hoppes, & Wilson, 1998; Jitendra, Hoppes, & Xin, 2000; Wong & Jones, 1982). One method of teaching main idea involves teaching students to identify the most important *who* or *what* in a section of text and then tell the most important information about that *who* or *what* in their own words (Jenkins et al., 1987; Malone & Mastropieri, 1991)

Using graphic organizers. Providing graphic and semantic organizers that assist students in writing or drawing relationships from the text has also been shown to improve students' comprehension. Among the strategies reviewed by the NRP (2000), teaching students to use graphic organizers was identified as having evidence of effectiveness. Using graphic organizers to enhance comprehension is particularly appropriate for expository text. When learning to use graphic organizers, students learn to visually represent the relationship among ideas in a piece of text. Commonly used graphic

organizers in content area classes include content webs to visually relate or compare topics and semantic features analysis charts, which allow students to compare features of different concepts, events or objects. Depicting these relationships supports students' comprehension and allows them to develop a more sophisticated model of the information presented.

In addition, there is evidence that using graphic organizers in science and social studies instruction benefits content area achievement (NRP, 2000). In a review of interventions for struggling readers, Mastropieri and colleagues (1996) found large effects (d = .92) for a construct defined as text enhancements, including graphic organizers. In a synthesis of interventions studying the effects of graphic organizers, Kim, Vaughn, Wanzek and Wei (2004) found an overall effect of close to one standard deviation.

Multiple strategies. Research shows that effectively teaching comprehension involves explicit strategy instruction of multiple strategies and opportunities for students to learn and apply the strategies in collaborative groups (NRP, 2000; Snow, 2002). Using multiple-strategy instruction can yield effective reading and comprehension outcomes across text types. In addition, there is evidence that for students with learning disabilities, multi-component strategy instruction combined with careful and gradual transfer to students is highly effective (Gersten, Fuchs, Williams, & Baker, 2001).

While multi-component interventions are effective, research indicates that teaching students a few strategies well (4-5) is more effective than teaching many strategies at a very superficial level. Among the most commonly studied multiple-strategy intervention, reciprocal teaching consistently yields moderate to high effect

sizes. For example, Rosenshine and Meister (1994) found large effects on researcher-developed measures (d = .87) and small to moderate effects on standardized measures (d = .36). Reciprocal teaching incorporates many of the strategies reviewed above, including question generation, summarization, clarification, and prediction.

Defining and documenting implementation: Implementation fidelity

Within the theory-driven evaluation model, an intervention's guiding principles provide the framework for understanding program outcomes. Examining the extent to which participants adhere to those principles and the related instructional practices enables the evaluator to attribute outcomes to what occurred rather than what was intended. Monitoring activities involve assessing not only the presence of essential program elements, but also the adequacy of implementation (Chen, 1990). Such data serve as a means of evaluating program outcomes within the context of real world implementation (Rossi, Freeman, & Lipsey, 1999).

Fidelity of implementation refers to the extent to which a program or intervention is implemented as intended across the entire intervention duration (Gersten, Baker, & Lloyd, 2000; Gersten, Fuchs, Compton, Coyne, Greenwood, & Innocenti, 2005).

Implementation fidelity data are often used to enhance internal validity by providing evidence that observed effects resulted from implementation of the treatment or program as intended. The WWC, a USDOE-funded project designed to cull from the literature effective practices in select domains, has identified treatment fidelity as a critical component in determining the validity of causal claims made in reported research (WWC, 2005). Treatment fidelity is one of the Evidence Standards used by the WWC reviewers to determine the strength of evidence from scientific studies in education.

Fidelity may also be assessed as a means of documenting variation within the context of real world implementation. In this case, information about treatment fidelity assists in understanding the relationship between critical intervention components and outcome data. When applied to program evaluation, fidelity of implementation serves as an indicator of the extent to which grantees or participants are implementing program components believed to influence program outcomes.

Fidelity measures are most often concerned with *whether* the treatment was implemented as planned. Such measures are typically used to address questions such as: (a) Was the implementation carried out for the required amount of time? (b) Were the required materials used? and (c) Were all the key features of the program (e.g., teacher modeling, teacher/student think-alouds, peer-groupings) implemented? (Mowbray et al., 2003). When conducting an experimental impact evaluation, implementation fidelity can also determine if critical elements of a program or intervention are absent in comparison classrooms (Kovaleski, 1999; Rossi, Freeman & Lipsey, 1999).

Defining and documenting implementation: Instructional quality

Fidelity measures also can be used to determine *implementation quality*. Complementing treatment-component measures, quality of implementation is concerned with how well an intervention was implemented. Judging quality can also occur at a more global level by rating features of instructional quality identified in the teacher effectiveness literature. Such features include student engagement, time on task, lesson pacing, and use of corrective feedback (Anderson, Everston & Brophy, 1979; Brophy, 1979). However, there is less known about the importance of documenting overall

instructional quality within a particular domain such as comprehension strategy instruction.

While the former indicator of fidelity, quantity, ascertains that the key features of the intervention/program were implemented as planned, the latter determines the extent to which the intervention's effectiveness is influenced by the quality of instruction or domain-related nuances believed to affect learning. Because quality indicators are usually high-inference variables, such measures are vulnerable to being less reliable (Shavelson, Webb & Burnstein, 1986). However, elements of instructional quality could potentially be critically important in terms of intervention outcomes.

Systems and methodological considerations for determining implementation fidelity

The most common method of obtaining implementation data is through classroom observations. An observation system is "a formalized set of rules for extracting information from the stream of behavior" (Hartmann & Wood, p. 108). The system includes how events are sampled, which behaviors or variables are targeted, and how they are scored or rated. Kennedy (1999) argued that classroom observations offer a first-level approximation of student learning, whereas teacher logs and teacher questionnaires would be second- and third-level approximations, respectively. In other words, the kind of instruction teachers provide—the instructional *content* of the class—offers the most direct indication of the knowledge and skills students are taught.

Process-product research on teacher effects has shown that opportunity to learn is a key predictor of student achievement (Brophy, 1979; Porter & Brophy, 1988). The process-product research defined opportunities to learn as a function of time allocated to active learning and content covered (Anderson, Everston & Brophy, 1979). Many of the

process-product observation studies focused on time allocated to active learning. In addition, studies of effective teachers of reading have consistently found that these teachers' classrooms are characterized by the relatively high amount of time students spend actively engaged in academic activities (Pressley, Rankin, & Yakoi, 1996; Taylor et. al, 2002). Rosenshine (1980), however, argued that of the variables related to opportunities to learn, content covered is more important than students' attention to the task or time engaged in academic activity because it is the most directly related to student outcomes. In other words, regardless of the time spent actively engaged in reading instruction, students will not perform well on outcome measures if they have not had the opportunity to learn the content being assessed. On the other hand, even if appropriate content is addressed, students will not have the opportunity to learn that material unless they are actively engaged in learning.

Because outcome measures may or may not measure knowledge and skills that are related to reading success, it is important to define content covered in terms of effective instruction rather than simply in terms of instructional alignment with assessments. In sum, opportunities to learn occur when instruction focuses on content and strategies that the research has shown to be related to student success in reading and when students are actively engaged in the instruction.

There are many observational systems from which to choose when designing a fidelity protocol. The desired response dimension—such as duration, frequency, or quality—can influence the type of system selected for a study. Whereas duration is an appropriate dimension when the proportion of time spent on an activity is of interest, frequency data are appropriate when the occurrence of an activity (regardless of time) is

of interest. Frequency data are also most appropriate when the target variables are of constant duration or when an individual is the unit of measurement (Good & Brophy, 2000). It is arguable that, based on these guidelines, frequency measures are desirable for studying implementation fidelity because program activities are prescribed and the teacher is the appropriate unit of measurement.

Depending on the research purpose, some systems are preferred over others. Common systems that yield frequency data include sign and time-interval systems. Researchers using a sign system observe behaviors for a pre-selected interval of time and then record all of the behaviors that occur during that interval (Borich, 2003). Because multiple codes can be recorded for each interval, categories do not have to be mutually exclusive or exhaustive.

A time-interval system, a very popular system, is similar to the sign system in that the observer records data at intermittent intervals of time. However, in a time-interval system the observer is required to choose a single category that best describes the behavior or activity during that interval (Borich, 2003). Therefore, the categories must be mutually exclusive. Time-interval systems are most appropriate for behaviors that occur fairly frequently, at least once every 15 minutes on average (Good & Brophy, 2000). In addition, creators of time-interval systems must decide whether a behavior (or activity) must occur during the whole interval to be recorded or whether occurrence during any part of the interval is sufficient to warrant coding. One of the biggest disadvantages with a time-interval system is that frequencies become a function of the interval length. Frequency counts are often converted to proportion of time variables by multiplying the frequencies by the interval length. When this is done, duration can be overestimated,

especially when long intervals are used and the targeted behaviors do not last the entire interval.

An event system is commonly used when duration, as opposed to frequency, is of interest. In an event system the observer waits for an event to occur and then records it (Borich, 2003; Good & Brophy, 2000). Event sampling is more appropriate when behaviors or variables are discrete or occur infrequently. The observer can either record an event as having occurred or he/she can record the initiation and termination of each event to yield data on event duration. While the latter approach appears to be most appropriate for measuring opportunities to learn in terms of time spent on effective instruction, real-time event recording is rather rigorous and can be difficult when the initiation and termination of the behavior or variables of interest are difficult to discriminate (Hartmann & Wood, 1990). However, this approach allows the researchers to record all instances of an event, instead of only those that occur during the sampled interval of time.

Dyadic interaction systems are used to code the interaction of the teacher with one student and are commonly divided into work-related contacts, procedural contacts, and behavioral contacts (Good & Brophy, 2000). Because the focus is on the teacher and a particular student, data on the behavior or activities for the rest of the class are lost. Therefore, such systems are not appropriate for the study of classroom-level variables such as instructional opportunities.

Alternative methodologies for measuring instructional content and format

Collecting implementation data through the use of trained observers, as done with many of the systems reviewed above, can be quite costly. Travel expenses, training costs,

and wages can consume a considerable portion of a project's budget, not the mention the time commitment required to conduct observations of a large sample. Power requirements in large-scale efficacy and effectiveness studies often dictate large samples of schools and classrooms (Raudenbush & Xiao-Feng, 2001), making observations a formidable and potentially "budget busting" task. In addition, although most federally funded research studies and program evaluations require an external evaluation, the budgets for such projects are often rather limited. In most evaluations and in many research projects, the ability to collect enough observation data to obtain stable, accurate measures of instructional practice will be constrained by available resources.

In an analysis of a set of reading instruction data, Rowan (2005) found that to obtain even minimally reliable estimates (r = .80) of certain key components in reading (e.g., comprehension and phonics instruction), a minimum of 10 observations per teacher would be needed. Even with a small sample of teachers, 10 observations of a 30-minute intervention can be resource intensive.

Of importance is identifying alternative, economical, yet effective data-collection practices to conducting classroom observations. Surveys, third-party observers, and logs are the most common. For example, several researchers interested in measuring the enacted curriculum have employed instructional logs (Camburn & Barnes, 2004; Porter, 2002; Smithson & Porter, 1994). Logs are most often developed by the researcher and completed by the teacher following the lesson or at the end of the school day. One researcher reported the cost of a log was approximately $28.00 (Rowan, 2005)—a figure well below the potential cost of a researcher-conducted classroom visit. Costs included in the above figure included training teachers to use the log, providing phone-based

support for answering questions, and providing a stipend for the time spent completing logs.

Figure 1 provides sample items from a log used by Camburn & Barnes (2004). The items on the log were developed to collect data on variables such as time on task, instructional focus, specific student learning tasks and teaching practices, and student engagement.

Figure 1. Sample teacher log items[1]

To what extent were the following topics addressed with the target student in reading/language arts today?

a. Text Comprehension:

◊ Primary focus ◊ Secondary Focus ◊ Touched on briefly ◊ Not a focus

What areas of comprehension did you work on with the student today?

- o Main idea
- o Summarization
- o Questioning
- o Previewing
- o Text Structure

1. Adapted from Camburn, E. & Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal, 105*, 49-73.

Results for reliability between trained observers and teachers differ in the literature. For example, the correlation between trained observers and teachers appears to differ across curricula domains, with higher correlations in studies of math and science instruction (e.g., Saxe, Gearhart & Seltzer, 1999) and somewhat smaller correlations in reading/language arts instruction (e.g., Rowan, Camburn & Correnti, 2004). This could be because reading/language arts instruction in grade 3 and above is often process oriented, with the "strands" or components often becoming indistinguishable and less transparent than a skill-based lesson typical in math instruction.

Camburn and Barnes (2004) examined the percentage of agreement on raters' reports of instructional practices and curriculum focus between researchers and teachers. They employed HGLM to determine how interrater reliability differed across the curriculum strands being measured and for different levels of detail (i.e., specific activities versus larger categories of instructional practice). When analyzing identical answers, these researchers achieved moderate reliability (r = .52) for teacher and researcher and for two researchers' (r = .66). When comparable answers instead of exact matches were assessed, reliability increased significantly to 81% and 87% respectively. In looking at reliability for individual components of early reading instruction, the areas of highest agreement were on activities related to comprehension, word analysis and writing (77-90% for teacher/researcher and 86-97% for researcher/researcher)—the components that were of primary interest to the researchers. In addition, at a grosser level of detail (e.g., was it taught vs. which activities were used or how much was it emphasized), reliability was considerably higher between teacher logs and observer records. This study also reported that disagreement between a teacher and a trained observer appears to be a function of teachers' background knowledge; teachers have exclusive knowledge of previous content, individual student needs and more to which observers do not have access.

The research on teacher logs tells us that given some concessions, an evaluator or researcher can reliably collect instructional practice data through this method. While the data may have to be collected at a more macro level than that which could be collected by a trained observer, the cost benefit may negate any loss of detail. Logs that request teachers to report retrospectively on instructional events or for the class as a whole often

distort the measure of opportunities to learn. Researchers often face decisions regarding sampling when using logs as well. For example, it must be decided if logs are completed for a day or week, a student or the whole class.

A review of the reading literature that reported details on implementation data collection revealed few instances of audio-tapes being used. Studies using audio-taped lessons had limited descriptions of how they were used. Roe and Vukelich (2001) used audio-tapes of representative lessons to assess fidelity of a reading tutoring program. Although the data revealed inconsistent fidelity, the data were not used to understand differences in program outcomes. Thus, little is known about the utility of this alternative methodology.

Observing reading instruction and analyzing instructional data

Within the reading intervention literature, it is still rare that instructional data are examined for their moderating effects on outcomes. Those studies that do are usually retrospective studies of the existing curriculum as opposed to studies of a specified program or intervention. For intervention studies that do assess fidelity, rarely is that data treated as an independent variable in modeling student outcomes.

Multiple studies that examined a variety of variables, including the nature of student/teacher interactions, elements of the classroom environment, and teacher behaviors, provided descriptive information on the instrument and the type of instruction provided, but did not analyze the relationship between instruction and outcomes.

Observation studies of the existing curriculum

Many instruments designed to capture opportunities to learn (i.e., instructional content variables) within the existing reading curriculum were found. In other words, the

following systems were not designed to assess fidelity to a specified instructional model or program. For example, the Multidimensional Reading Instruction Observation Scale (McCabe, 1992) is a descriptive framework used to describe cognitive and affective teacher and learner processes and teacher management skills during reading instruction. The GRIP (Magano, 1982) captures patterns of verbal behavior (e.g., positive reinforcement) and the teaching/learning process in the reading classroom through the use of a time-interval counting system.

The Reading Lesson Observation Framework (RLOF; Henck, Moore, Marinak, & Tomasetti, 2000) was developed as a tool for supervisors to use in evaluating teacher behaviors during reading. The tool consists of seven domains—classroom climate, prereading phase, guided reading phase, postreading phase, skill and strategy instruction, materials and tasks of the lesson, and teacher practices. Each domain is further defined by 5-11 indicators. Observers make judgments about each item, scoring it as either observed/high quality, observed/low quality, not observed, or not applicable.

While studies employing the above instruments provided only descriptive information on instruction, Foorman and Schatschneider (2003) developed a system that has been used to collect data intended to serve as an independent variable. They combine a time-sampling and checklist approach to observing reading instruction. The system developed by the researchers involves four components: (1) a time-by-activity measure, (2) a student engagement measure, (3) a teaching strategies checklist, and (4) a teacher effectiveness rating scale. For the first 10 seconds of each instructional minute, the observer codes the instructional format (e.g., grouping pattern) and content of teaching. A

checklist of specific instructional strategies is also completed at the conclusion of each observation. Items on the checklist are marked as either observed or not observed.

While the categories used are aligned with the research, the time-sampling approach employed presents several problems. First, reading instruction is complex and, often, the instructional objective of an activity may not be apparent until the activity has been observed in its entirety. For example, what first appears to be a teacher read aloud may in fact be a structured oral language activity during which the teacher uses the text to facilitate children's oral language development. With a time-interval system such as this one, the observer is forced to make a decision about the instructional content without seeing the activity as a whole event. When observers are unable to make a quick determination of the instructional content, minutes are "dropped" from the observation and data are lost. Lastly, because one category is selected during each interval, if more than one instructional event is occurring (e.g., several centers are operating simultaneously) only one component will be coded.

Observers also rate the level of engagement for four randomly selected students at 10-second intervals. To measure teacher quality, a checklist of teacher competencies is completed and an overall global rating of teacher quality is assigned. In subsequent research using this system, the correlation between the quality checklist and overall global rating of teacher effectiveness was found to be sufficient ($r = .63$) to warrant using only the overall global rating as an indicator of teacher quality (Foorman, Schatschneider, Eakin, Fletcher, Moats and Francis, 2006).

This overall rating of teacher effectiveness was entered in a multi-level model that included the following predictors: initial reading status, grade, and time spent on

components of reading instruction. Results indicated that initial reading status was a moderate predictor of end-of-year reading outcomes (Foorman, Schatschneider, Eakin, Fletcher, Moats and Francis, 2006). There was only one main effect for time allocated: time spent engaged in reading versus giving directions explained significant variance in the reading outcomes. Although comprehension was one of the time allocation categories, that variable accounted for approximately 4% of the between-level variance.

Interestingly, there were significant, but weak, effects of teacher quality ratings on all reading outcomes. In several applications of this observation system using the overall rating of teacher effectiveness, good teaching was found to make a small, but sometimes significant, difference in improving reading comprehension beyond initial reading abilities. It is important to note that the reading curriculum in its entirety was examined and other studies have found that higher-level comprehension instruction is often a very small percentage of the time allocated to reading instruction (Connor, Morrison, & Petrella, 2004; Durkin, 1979; Pressley, 1998).

Taylor et al. (2002) developed the School Change Observation Scheme for Classroom Literacy Instruction (SCOS) as a tool to collect data related to grouping practices, literacy events, materials , interaction styles, expected student responses, and students' engagement rate during reading instruction. Observers using the SCOS, a sign system, sample instruction every five minutes, write a narrative of what is observed, and then code the instruction that occurred during the interval along the levels listed above. Applications of this system to the study of instructional implementation have shown that teachers in highly effective schools spend more time on higher-order skills such as comprehension.

Similar to the overall teacher effectiveness rating used in the Foorman system, Taylor and colleagues developed a teacher accomplishment rating, which is an overall rating of teacher accomplishment based on the presence of such characteristics as teacher enthusiasm, task orientation, and student engagement. Teachers were rated on "accomplishment" by experienced educators on a scale of 1 (Least) to 3 (Most). Ratings of teacher accomplishment were positively correlated with several other variables including fluency scores, time on task, small group work, and using a coaching/scaffolding approach to teaching.

In subsequent research using this system, Taylor, Pearson, Peterson and Rodriguez (2003) used HLM to show that a number of teaching variables explained variance in student reading growth. For example, teachers who emphasized higher order thinking showed greater growth in their classroom's reading achievement. In Grades 2-5, 48% of the variance in comprehension scores was between teachers, after accounting for pretest scores. When that variance was modeled as a function of instructional practices, 20% of the between teacher variance was accounted for by the following variables: higher order questioning, time on task, teaching comprehension skills over comprehension strategies (negative relationship), and passive responding (negative relationship). Qualitative analysis of teachers with high scores on higher order questioning revealed that these teachers engaged in such practices as having students work in pairs to ask/answer questions about text, summarize and retell, make predictions, and preview.

Although the system of determining instructional opportunities was not fully explicated, another study of the enacted reading curriculum for 2nd-5th grade students

found a positive effect of teacher-led comprehension instruction on reading outcomes (Conner, Morrison, & Petrella, 2004). In this study, instruction, as opposed to student characteristics, explained 33% of variance on post-test comprehension outcomes (Connor, Morrison, & Petrella, 2004). An interaction between initial reading skill and the type of instruction provided was found, with lower to average readers benefiting more from teacher-led comprehension instruction. It should be noted, however, that of all the components addressed during the observer reading instruction, actual comprehension strategy instruction lasted an average of only minutes per day, suggesting that greater doses of comprehension instruction could positively impact student learning.

Intervention studies with implementation fidelity assessed

The systems and outcomes described above all focused on the overall, existing reading curriculum. Although it is increasingly more common to assess fidelity of implementation while conducting intervention research studies, it is less common to examine the impact of implementation on results. This is likely because in most intervention research high levels of fidelity are maintained to eliminate variation as a possible confound. As the WWC guidelines suggest, only in instances where treatment variation is believed to occur, such as a real-world program evaluation, would implementation data warrant attention. As recently as 1999, a review of the phonological awareness literature, a large body of work in the reading field, found "insufficient or nonexistent assurance of fidelity to treatment" (Troia, 1999). In the reviewed studies that did assess fidelity, few examined the relationship between implementation and outcomes, not all studied reading, and those that did often focused on reading components other than comprehension.

For example, Fuchs, Fuchs and Karns (2001) assessed fidelity to a math intervention in the elementary grades. In a practice that appears to be quite typical in assessing fidelity, intervention components were rated as observed, not observed, or not applicable (Fuchs, Fuchs and Karns, 2001). The researchers included both teacher components and student components on their fidelity instrument because peer-assisted learning was an integral part of the intervention.

The researchers derived a "percentage implemented" score by dividing the number of observed behaviors by the number observed plus not observed. Such procedures provide a consistent indicator of implementation even when lessons vary in the number of components included from day to day. For example, teacher modeling may only occur in 1 lesson out of every 3, but under this system the teacher would not be penalized for omitting this practice when appropriate. However, despite this detailed system, the data were not used as explanatory variables.

Jackson, Paratore, Chard and Garnick (1999) conducted observations of an early reading intervention using a checklist to rate behaviors as either *present* or *absent.* In addition, teachers completed weekly logs containing items aligned with the checklist. Although the fidelity data from this study were not analyzed quantitatively, a review of field notes indicated that there was a great deal of implementation variation for program behaviors that had been encouraged but not required. Mathes, Torgesen and Allor (2001) reported high levels of fidelity when teachers implemented a phonological awareness intervention. However, other than reporting that implementation levels were high, no data were presented or analyzed.

In one study of a specific reading program, higher levels of fidelity resulted in larger gains in literacy scores (Frechtling, Zhang & Silversteing, 2006). In this study, researchers used an instructional fidelity index protocol, rating each component on a 4-point scale ranging from 0 (not implemented) to 3 (implemented effectively). Ratings indicated the quality and extent to which components were implemented (e.g., the teacher has established independent reading stations and uses them appropriately). A total implementation score was based on the sum of ratings across all items. A multi-level model regressed gain scores on implementation, percent male, class size, teacher experience, and percent free and reduced lunch. Two analyses were conducted, one using discrete scores and one ordinal scores. When ordinal scores were used, the level of implementation had a statistically significant effect on student achievement on seven measures of reading ($p \leq .05$).

In sum, assessing fidelity is commonly done through rating the presence of specified behaviors, but resulting data are rarely examined in terms of their impact on outcomes. Few intervention studies examine the relationship between fidelity and outcomes (Foorman, Chen, Carlson, Moats, Francis, & Fletcher, 2003; Frechtling, Zhang, & Silverstein, 2006). Although some studies of early reading skills reported high levels of fidelity for teacher-implemented interventions, this is likely not the case for reading components such as comprehension that require highly skillful teaching. However, because studies examining implementation fidelity in comprehension interventions is largely absent from the field, the question remains to be answered.

Chapter 3: Methods

Context and participants

Context. Elementary schools from two urban districts participated in the study. Three schools from one district and 2 from the other district (5 schools total) were randomized to the comprehension condition and are included in the study. Both districts were comprised of diverse student populations and had high percentages of students qualifying for free and reduced lunch programs. A randomized cluster design was utilized with schools matched on demographics before being randomly assigned to condition.

Teachers. The study examined data on 4th grade social studies teachers implementing a research-based comprehension intervention program (n = 16). Table 1 provides teacher characteristics. All teachers except one were certified in elementary education. Teachers had an average of 9 years teaching experience (M = 9.62, SD = 9.93) and an average of 5 years in 4th grade (M = 5.15, SD = 5.89).

Table 1
*Teacher characteristics**

	Percentage (%)
Female	85
Ethnicity	
African American	31
Hispanic	23
Caucasian	46
Highest degree earned	
Bachelors	92
Masters	8
Additional certifications	
ESL	39
Bilingual	31
Other	23
Bilingual teacher	23
Departmentalized classrooms	77

* $n = 13$

Students. Data were collected from 309 4th grade students from participating teachers' classrooms. Table 2 provides information on student characteristics.

Table 2

Student characteristics

	*N**	Percentage (%)
Gender	294	
Female		51
Male		49
Ethnicity	290	
African -American		21
Hispanic		65
White/Non-Hispanic		13
Free/Reduced Lunch Plan	296	71
Limited English Proficiency	294	25
English as a Second Language	290	19
Qualified for Bilingual Education	294	5
Qualified for Special Education	269	1

* Student characteristic data were not available for all students. *N* = number of students for whom data were available for each characteristic.

Participant consent. The study is a sub-analysis of extant data from a study with Institutional Review Board approval (IRB protocol #2005040097). No additional data collection was conducted. Consent to participate in the larger study was obtained from all teachers included in the sub-analyses. Only students with parental consent and student

assent were administered assessments and were included in data analysis. Copies of all consent forms can be found in Appendix A. IRB approval was obtained to conduct this secondary analysis (IRB protocol # 2007-05-0006).

Comprehension program description

The comprehension intervention examined in this study was one part of a multi-component program that included both comprehension and vocabulary instructional practices. Because schools were randomly assigned to either the comprehension or vocabulary group, sub-analyses of the separate practices are possible. The description below focuses only on the comprehension practices. The program consisted of three modules, each featuring training, lessons and materials for teaching students a new comprehension strategy. Teachers agreed to implement lessons from each module three times a week for 30-minutes over a 6-week period during social studies class, (a total of 18 lessons per module, 54 lessons total). Module content was additive. That is, strategies learned and applied from the first module were continued and added to with additional strategies in the second module.

Purpose. The overall purpose of the professional development program was to enhance teachers' knowledge of instructional practices in reading comprehension that promote academic success in the content areas, specifically social studies. Despite the substantial research base supporting the efficacy of explicit comprehension strategy instruction and the availability of information and tools for providing high-quality comprehension instruction, such practices have not found their way into many classrooms (Pressley, 1998; Snow, 2002). Thus, a primary purpose was to bridge the gap between research to practice and provide the incentive for teachers to implement the practices in

the form of high quality professional development and structured lessons connected to the content.

The program's professional development sessions provided 4th grade teachers opportunities to learn research-based instructional strategies. Participating teachers were charged with teaching students how and when to use specific comprehension strategies flexibly and in combination. Although specific strategies were taught, the overall objective was for students monitor their understanding (asking/answering questions and writing gist statements and summaries) and independently apply these comprehension monitoring strategies while reading expository text.

An explicit comprehension instructional framework was used to teach the strategies. Multiple studies and syntheses have demonstrated the effectiveness of teaching comprehension strategies directly to students, especially struggling readers (Duffy & Roehler, 1982; Gersten, Fuchs, Williams, & Baker, 2001; NRP, 2000; Pearson & Dole, 1987; Swanson, 1999). Hallmarks of explicit strategy instruction include a direct explanation of the strategy with teacher modeling followed by both guided and independent practice. The approach supports the gradual release of control from the teacher to student through initial teacher modeling (I do it), teacher-assisted practice (we do it), and independent student practice (you do it).

The research team selected social studies because comprehension of content-area texts can be challenging for upper elementary grade students (Jetton & Alexander, 2004; Juel, 1998). In addition, social studies was selected over science because social studies instruction is more often text based and the two participating districts had adopted the

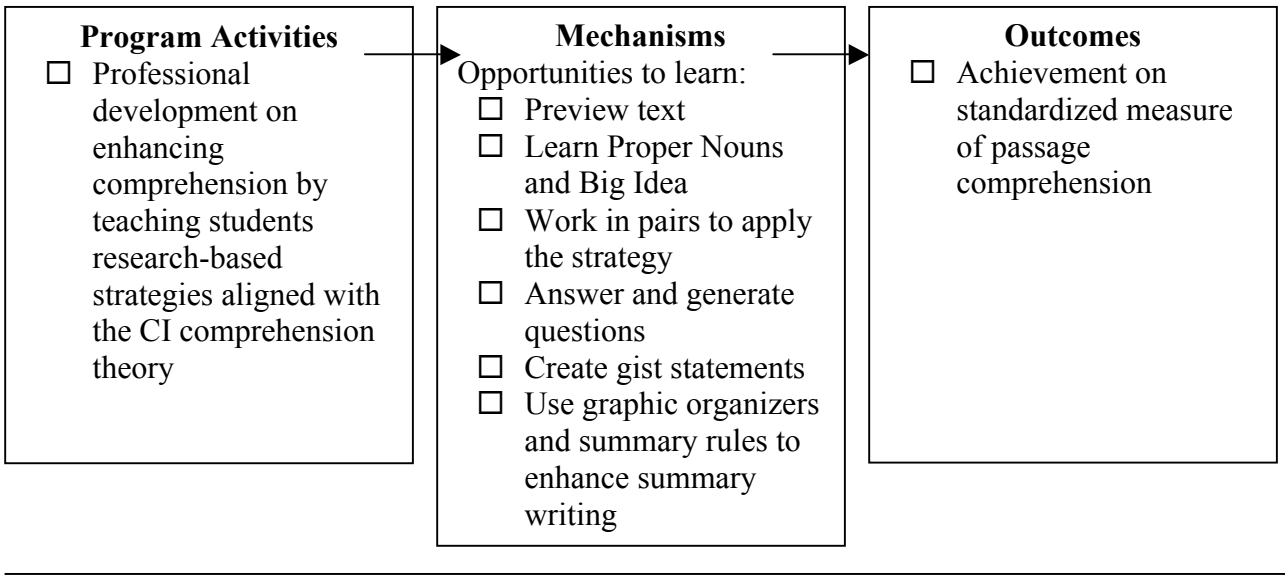
same social studies curriculum (Social Studies Texas, Scott Foresman 2003), thereby eliminating content as a potential confounding factor.

Because the strategies included in the lessons consistently yield high effects in the intervention literature, this was not an effectiveness study. Rather, the intent was to examine the effects of offering a professional development program with information and materials on best practices in comprehension and asking teachers to implement the lessons in the real-world context of a 4th grade social studies class.

The theoretical model that guided the development and evaluation of the comprehension program is illustrated in Figure 2. The program's primary goal was to enhance comprehension by teaching students to improve their comprehension of expository text information through strategy instruction. Program activities consisted of lessons, materials and training on effective comprehension practices. These are described in more detail below.

As seen in the program model, the mechanism by which these activities affect comprehension is through increased instructional opportunities. Instructional opportunities serve as the mechanism of change because such opportunities are readily observable and measurable. Although one could argue that the mechanism by which program activities result in the intended outcome lies within a student's cognitive and metacognitive skills (e.g., developing mental models of novel information in text), such variables are not easily measured and when they are it is often through resource-intensive student interviews, which is cost-prohibitive when conducting analyses using large-sample methods. In this instance, opportunities to learn occurred when instruction focused on the strategies featured in the professional development.

Figure 2. Theoretical program model for comprehension intervention



Development. Lessons and materials were developed during the 2005-2006 school year by a team of researchers with both classroom and research experience in the area of comprehension instruction. The team culled the literature on effective comprehension practices and identified those that consistently yielded high effects and were most appropriate for content area instruction. Experts in the reading field reviewed lesson content and format and provided feedback.

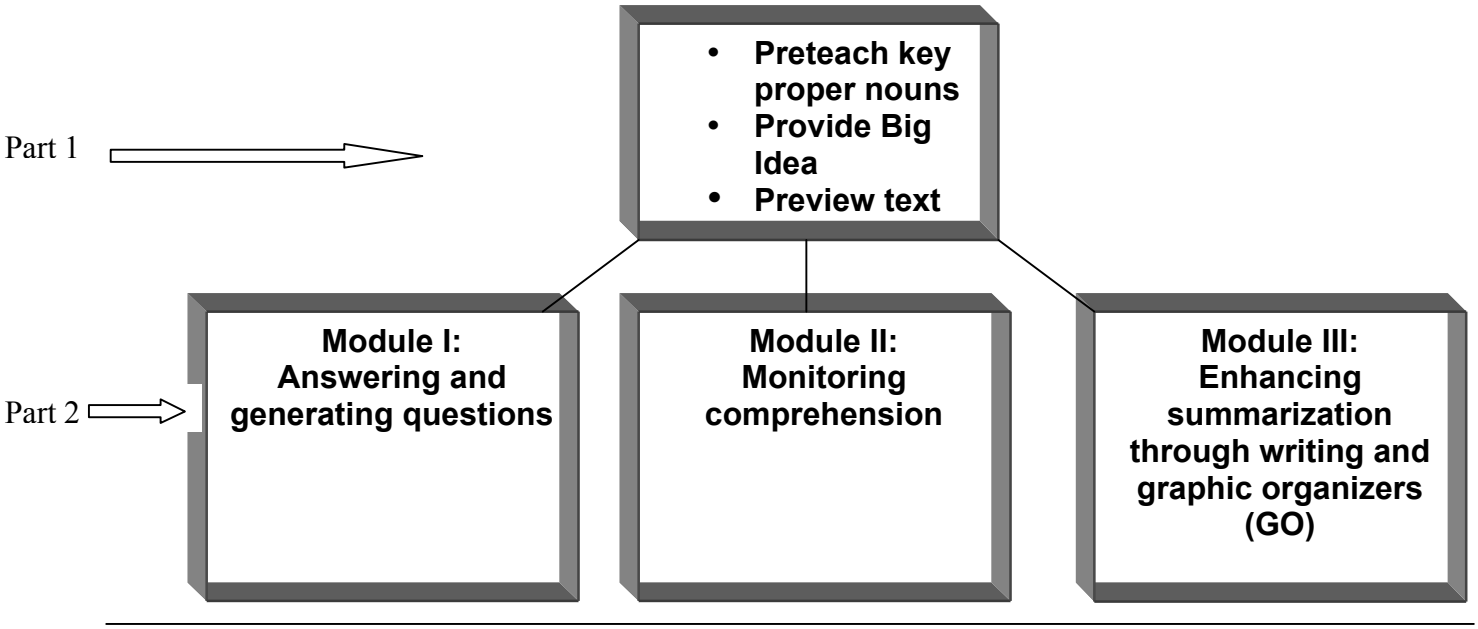
In addition, three veteran teachers in a local school district piloted the lessons and provided formative feedback, focusing primarily on feasibility within the real world of a 4th grade classroom. The development team revised lessons to address the suggestions gathered during the pilot. Upon finalizing lesson content and materials, the team developed professional development materials and activities.

Comprehension program components. As mentioned previously, the program consisted of three modules, with each module introducing a new strategy and building

upon previously learned strategies. The modular design was intended to scaffold instruction by introducing one element at a time and providing students with multiple opportunities to practice before introducing the next step.

Figure 3 illustrates the comprehension program components. Part 1 includes instructional practices featured in every lesson. These practices include three pre-reading activities: (a) preteaching or reviewing proper nouns; (b) previewing the text; and (c) providing the Big Idea for a section of text. Part 2 in Figure 3 shows the strategy introduced in each module. In addition, the lessons were taught using an explicit instruction model that included a teacher modeling phase, a guided practice phase and an independent practice phase. During the guided and independent practice phase, students worked in Student Study Teams, an adapted classwide peer tutoring model that incorporates peer tutoring to support student learning (Fuchs, Fuchs, Mathes, & Simmons, 1997; Greenwood & Delquardi, 1995; Fulk & King, 2001).

Figure 3. Comprehension intervention program components.



Module 1 introduced a questioning strategy adapted from Question and Answer Relationships (QAR; Raphael, 1986; Raphael, Highfield, & Au, 2006). Through scaffolded, explicit strategy instruction, students were taught to ask and answer three different question types:

- *Level 1: Right There*. Answers are explicitly stated, word for word, in one place in the text.
- *Level 2: Putting it together*. Can be answered by looking in the text, but require the reader to put information together from different parts of the text.
- *Level 3: Making Connections*. Cannot be answered by looking in the text alone; require students to think about what they have just read, what they already know, and how it fits together.

Students learned to Get the Gist, or main idea, in Module 2. Teachers scaffolded instruction by first modeling how to:

1. identify the most important who or what in a paragraph and
2. tell the most important thing about that who or what in 10 words or less.

As students gained proficiency in writing gists at the paragraph level, larger units of text were assigned. Once students understood how to generate a main idea statement, they incorporated Get the Gist with the questioning strategy learned in Module 1. Students were instructed to ask themselves questions after reading a section of text to help them monitor their understanding and generate a gist statement.

In Module 3, students learned to summarize larger pieces of text by utilizing a content web. Again, strategies from the first two modules were incorporated into Module 3 lessons. For example, students organized their gist statements on the content web and then used the relationships depicted on the web to write a high-quality summary. Lessons included guidelines to help students develop a summary from their main idea statements. These summary rules included: (a) write a topic sentence using the Big Idea, (b) include gists, (c) delete information that is redundant or trivial and (d) re-read to make sure the summary makes sense.

Lesson format. Although lessons varied depending on where they fell within the explicit instruction cycle (i.e., explain, model, guide), there was a standard format. Figure 4 presents the typical lesson format. Each lesson was designed to last 30 minutes, with 5 minutes allocated for the pre-reading activities (steps 1-3). At the beginning of each lesson, the teacher pre-taught proper nouns and provided the Big Idea for a new section of text. If the class was continuing with a previously begun section, key proper nouns and the Big Idea were reviewed. For each new text section, the teacher led a preview.

Initial lessons in each module included a teacher model phase, followed by a guided practice phase and independent practice phase. A sample lesson sequence for Module 1, showing the gradual introduction and transfer of a new practice from teacher to student, can be found in Appendix B.

Figure 4. Lesson format for comprehension intervention

1. Preteach/review key Proper Nouns (teacher)
2. Introduce the Big Idea (teacher)
3. Preview the passage (teacher and students)
    - √ What I already know
    - √ Make a prediction
4. Model/guided practice/review strategy (teacher/students)
5. Read with a partner (student pairs)
6. Apply strategy (student pairs)
7. Share product with feedback (teacher/students)

Materials. Teachers were provided with all necessary materials to implement the lessons. Each lesson contained an objective, list of materials, instructions for teacher preparation and a scripted lesson. Planning sheets provided space for teachers to write sample questions, key proper nouns, etc. Teachers were not required to read the script, but it was provided as a support if needed. Teacher notebooks included transparencies to support instruction and application. For example, a transparency on how to work in student pairs listed the rules for working cooperatively with a partner.

Student materials included a folder with logs for documenting progress and providing evidence of understanding. A sample student learning log is provided in Appendix C. In addition, cue cards for each strategy served as a reference for students as they were learning new strategies. Sample cue cards are included in Appendix D.

Training. Teachers participated in three half-day training sessions, one for each Module. Training consisted of an overview of the featured strategy, explanations of lessons and materials, and opportunities to practice applying the strategies (i.e., role play activities). In addition, teachers attended three 90-minute teacher study team meetings to discuss implementation and problem-solve with the research team.

Measures

Student Outcomes. The primary student outcome, comprehension achievement, was measured using the passage comprehension subtest of the Gates MacGinitie Reading Tests 4th edition (GMRT, MacGinitie, MacGinitie & Dreyer, 2000). Students were assessed one to two weeks before instruction began to establish a baseline and within 2 weeks following the intervention. The GMRT is a 35-minute, group-administered standardized assessment of comprehension achievement with 11 passages and 48 multiple-choice questions. Items consist of both inferential and literal multiple-choice questions. Reliability coefficients (KR-20) of .93 and .92 for the comprehension subtest at Level 4 meet research standards. Trained testers administered both pre- and post-tests to consented students.

Research has documented relatively weaker effects on standardized measures compared to researcher-developed measures, which tend to be more closely aligned with the intended program (e.g., Rosenshine & Meister, 1994). A sub-set of students was

administered a post-intervention strategy-use interview developed by the research team to assess process information (strategy knowledge and use). However, because the sample interviewed was so small (n = 16), this more proximal measure of program outcomes is not included in the analyses. Using a standardized measure of reading comprehension offers a more rigorous test of implementation effects on outcomes.

Implementation fidelity and teacher quality. We documented implementation in two ways. First, teachers completed weekly instructional logs. Logs provided both an indication of dosage—the number of lessons implemented in a given week—and implementation fidelity in terms of components taught. For each of 5 components, teachers indicated which had been taught for each lesson that week. Appendix E features a sample teacher log. The average length of a lesson and a weighted and unweighted sum of components implemented will be calculated for each teacher for select log submissions.

Secondly, each teacher recorded a representative lesson for each Module using a digital recorder. Insufficient staff resources prohibited classroom observations, thereby necessitating the use of recorded lessons. Lesson files were uploaded from the recorder and coded by members of the research team using the fidelity protocol.

The two research teams at TAMU and The University of Texas collaborated to develop the fidelity protocol. The coding protocol was adapted from similar instruments used in other research studies (Vaughn, 2002; Vaughn, 2001). The adapted instrument included instructional variables representing the research-supported comprehension components (e.g., preteaching key proper nouns, teaching Get the Gist) and variables describing the elements of explicit strategy instruction (e.g., modeling, guided practice).

Critical intervention components were selected for inclusion based on an empirical basis and the hypothesized relationship to outcomes. In keeping with the literature on reliability of observation data, macro-level items were retained and specific items about procedures and interactions were eliminated.

The final protocol included 8 items (see Appendix F). Six items represent implementation of the major program components and are rated on a 4-point scale ranging from 0 (not at all) to 3 (exemplary implementation). Rating items using a Likert-type scale allows one to measure both the presence/absence of components and the *level* of implementation simultaneously (e.g., acceptable, exemplary). The research team realized that teachers would likely modify the practices somewhat to better fit their students' needs. The team agreed that modifications should not be so great that they deviate substantially from the original model or significantly alter key components. Therefore, indicators for each codeable item were developed to provide a priori guidelines for determining acceptable implementation (see also Appendix F). Although the ratings require some level of inference, providing explicit guidelines and using a small number of coders who are intimately familiar with the intervention supported optimum reliability.

In addition to items related to specific components implemented, two additional summary items were included. These items were adapted from existing observation systems (Foorman & Schatschneider, 2003; Foorman et al., 2006; Taylor et al., 2002; Taylor et al., 2003). For Item 7, coders rated the overall program implementation on a 7-point semantic differential scale ranging from 1 (less than adequate) to 7 (above expectations). Raters used the ratings on items 1-6 to inform their rating on item 7.

An item of overall teacher quality, Item 8, was used to provide an overall assessment of a teacher's instructional quality. Teacher quality was rated on a 7-point semantic differential scale, ranging from 1 (not at all effective) to 7 (highly effective). The inclusion of this item was based on the work of others who have found that a single item rating overall instructional quality explains significant variance in student outcomes (e.g., Foorman et al., 2006; Taylor et al., 2002; Taylor et al., 2003). All raters had both classroom experience and extensive experience observing teachers, which enabled them to make a judgement about overall instructional quality. Because the number of raters was small (3 raters were used), the team was able to have a thorough discussion about what constituted high quality during initial training. During the discussion, elements such as lesson pacing, student monitoring, corrective feedback, clarity and explicitness of explanations, and scaffolded instruction were identified as potential indicators of high quality instruction.

A three-member team coded 2 randomly sampled recordings of comprehension instruction per teacher (approximately 40 recordings). All 3 recordings were coded for 10% of the sample (n = 2). The random selection was counterbalanced such that for a third of teachers, the first two recordings were coded, for another third, the second two recordings were coded and for the last third, the first and last recording were coded. After initial training, each member coded two randomly selected tapes. Coders demonstrated 88% agreement prior to coding independently. Agreement was calculated as the number of agreements divided by the number of agreements plus disagreements across all 8 items. In addition, 20% of recorded lessons will be double-coded and assessed for inter-

rater reliability. In the event of a discrepancy between coders, items in question were reviewed and rated by an expert coder.

Research Questions

Research Question 1

Do classrooms vary in comprehension achievement?

Hypothesis 1. There will be significant between-classroom variation in post-test comprehension scores.

Rationale 1. Classroom intervention research has found as much as 25%-50% of variance in student outcomes to be between classrooms (Frank, 1998).

Methods 1. Hierarchical linear modeling (HLM) was used to answer Questions 1-3. Analyses was conducted using the statistical software package HLM 6 (Raudenbush, Bryk, & Congdon, 2006). HLM, which allows the analyst to include additional sources of variance in modeling student outcomes, is an appropriate methodology because of the data's nested structure. A two-level model was examined with students, the level-1 unit of analysis, nested within teachers, the level-2 unit of analysis. The importance of considering teacher- and student-level characteristics simultaneously within a multi-level model has been well documented (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Using traditional methods that do not account for the effects of group membership would result in smaller and biased standard errors, thereby inflating test statistics for model coefficients and increasing the Type-I error rate.

To answer Question 1, an unconditional 2-level model was used:

Level 1 $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2 $\beta_{0j} = \gamma_{00} + u_{0j}$

where Y_{ij} is an individual student's comprehension score at post-test. Specifically, to determine the presence of significant unexplained variance between classrooms, I tested the null hypothesis:

$$H_0 : \tau_{00} = 0$$

While the significance of τ_{00} provided a direct answer to Question 1, the intra-class correlation is also reported as a descriptive index of the proportion of total variance that is between classrooms. In addition, this model provided an overall average post-test comprehension score.

Research Question 2

To what extent do ratings of implementation fidelity and overall teacher quality affect student comprehension outcomes after controlling for initial reading status?

Hypothesis 2. Ratings of implementation fidelity and overall teacher quality will explain significant variability in comprehension outcomes above and beyond students' initial comprehension skills.

Rationale 2. Overall ratings of teacher quality has accounted for significant variability in reading-related outcomes (e.g., Foorman, Francis, Fletcher, Schatschneider & Mehta, 1998; Taylor et al., 2003), most often after controlling for initial reading status. Many intervention studies that have examined implementation fidelity have focused on early reading skills such as phonics and phonemic awareness, so little is known about the impact of variation in fidelity for programs addressing the more complex skill of comprehension. A positive relationship between time allocated to higher order skills within a reading curriculum and student outcomes has been demonstrated (e.g., Connor, Morrison, & Petrella, 2004; Taylor, Pearson, Clark & Walpole, 2000), yet there is little

research on the predictive power of fidelity to a comprehension intervention program, so this hypothesis is somewhat exploratory. In previous studies, a positive relationship to achievement was found with only minimal amounts of comprehension instruction within the reading curriculum. Thus, implementation ratings from a 30-minute comprehension program like the one being considered in this study is likely be positively related to outcomes.

Research Question 3

Is there an interaction between implementation fidelity and teacher quality on student comprehension outcomes after controlling for initial reading status?

Hypothesis 3. Ratings of implementation fidelity and teacher effectiveness will interact to create differential effects on outcomes.

Rationale 3. Research has shown a positive correlation between teacher quality and the presence of instruction on higher-level skills, with highly effective teachers spending more time on higher-order skills in reading (Pressley, 1998; Connor, Morrison & Petrella, 2004). In addition, a long line of research has demonstrated the association between effective teaching and student achievement (see Brophy, 1979). Although little research has examined the effects of differing fidelity on outcomes, the high effects for the strategies featured in this program suggest that faithful implementation could yield higher effects despite poor quality teaching overall. The teacher effectiveness literature, in combination with the strategy intervention research, suggests that an interaction between implementation and teacher effectiveness will exist. More specifically the data suggest that at higher levels of implementation, the effects of teacher quality will be

diminished such that achievement between classes with high and lower quality teachers will be less pronounced.

Methods 2 and 3. To address Questions 2 and 3, a contextual analysis was conducted to determine the impact of group level predictor variables (fidelity and teacher effectiveness ratings) on within-group coefficients. More specifically, a fully conditional 2-level model was used to examine the moderating effects of fidelity ratings and overall teacher effectiveness ratings on students' comprehension scores. By examining teacher-level predictor variables, I determined whether students from different classrooms show systematic differences in (a) comprehension achievement and (b) the strength of relationship between initial reading status and end-of-year comprehension scores.

The model building process proceeded as follows. A random-coefficient (RC) model was tested to provide a baseline against which to compare a fully conditional model. At level-1, individual comprehension post-test scores were modeled as a function of the sample average and the student's group-mean centered pre-test score (i.e., the student's deviation from the class-average pre-test score). The RC model tested was:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\mathrm{Pr}e_{ij} - \overline{\mathrm{Pr}e}_{j}) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

where Y_{ij} is an individual student's post-test GMRT score and *Pre* is the GMRT pre-test score. In this model, β_{0j} is the expected class average and β_{1j} is the expected pre/post slope. Variance for each level-2 outcome was tested using a chi-square test of significance. Significant variance of the slope and intercept provided guidance for the next stage of model building (e.g., modeling unexplained variance in slope or intercept

across classrooms after controlling for pre-test). However, even though one variance estimate was not significant, model building proceeded to test explanatory variables at Level 2.

To model variation in student outcomes, two explanatory variables—teacher-level ratings of fidelity and quality—were tested with a fully-conditional model. The level-1 model remained the same. The new level-2 model was:

$$\beta_{01j} = \gamma_{00} + \gamma_{01}(FID_j) + \gamma_{02}(TE_j) + \gamma_{03}(PRE_j) + \gamma_{04}(FID_j, TE_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(FID_j) + \gamma_{12}(TE_j) + \gamma_{13}(PRE_j) + \gamma_{14}(FID_j, TE_j) + u_{1j}$$

where *FID* is an average fidelity of implementation rating, *TE* is an average overall teacher quality rating across all coded recordings, and PRE is the class average pretest score on the GMRT. The two implementation variables were grand mean centered to assist with interpretation. Results are reported for the fully conditional model. In addition, any of the classroom-level predictor variables that were not significant at the $p = .10$ level were deleted from the model. A final model with non-significant level-2 explanatory variables deleted was tested.

Reported results will include:

- Significance test results for the effects of interest (γ_{10} , γ_{01}, γ_{02})
- Significance test results for the interactions of interest (γ_{11}, γ_{12}, γ_{04}, γ_{14})

Significant interactions are depicted graphically to assist with interpretation.

Discussion and follow-up analyses examine:

  - Comprehension achievement (student outcomes in classrooms with teachers who are below and above the mean on implementation fidelity and overall quality).

- Equity in instructional impact (pre/post slopes in classrooms with teachers who are below and above the mean on implementation fidelity and overall quality).

Research Question 4

What is the correlation between teacher logs and research-coded protocols on length of session (dosage estimate) and number of components implemented for sample lessons?

Hypothesis 4. There will be a significant, but moderate positive correlation between the two sources of information on both length of session and the number of components implemented.

Rationale 4. Research indicates moderate to high correlations between researcher codes and teacher logs when the data requested are at a macro level (Camburn & Burns, 2004; Smithson & Porter, 1994). In addition, correlations between teacher reports on comprehension practices and researcher codes has been reported as high as $r = .9$ (e.g., Camburn & Burns, 2004). It should be noted that in the study cited above, 18 instructional practices were included in the comprehension construct. The intervention in this study includes only a small number of comprehension practices.

Methods 4. Pearson correlations were calculated and tested for each variable of interest: average length of session and number of components implemented (average fidelity score). Although it is not possible to correlate data from the same lesson (teachers did not often specify which lesson was recorded), it is possible to compute an average length of session and components implemented for the two-week window in which the recording took place. Because lesson components are not equally important, a weighted

raw score was computed such that more important components were weighted more heavily ($w = 3$) in the computation of a raw total score. There are 3 lesser components (previewing, pre-teaching proper nouns and providing the big idea) and 2 major components (modeling/guided strategy instruction and independent student strategy application). Thus raw scores ranged from 0 to 9, with 9 indicating all components were implemented. The teacher-provided data was correlated with researcher data gleaned from the recorded lesson. In addition, to support the subjective decisions about weighting select components, an unweighted raw score also were entered into analyses to see if that in fact changes the results. Results from both the unweighted and weighted analyses are reported

Research Question 5

Are estimates of teachers' implementation stable across occasions?

Hypothesis 5. Two fidelity checks will be sufficient to obtain stable estimates of teachers' implementation.

Rationale 5. The literature on observation research varies considerably in recommendations for ideal number of observation occasions. Research both past and current suggest a minimum of 3 to 5 and up to 10 observations are needed (depending on the component or behavior of interest) to obtain a stable estimate of instructional opportunities present in a teacher's classroom (Rowan, 2005; Erlich & Shavelson, 1978; Shavelson & Dempsey, 1976). However, in this literature, researchers observed the enacted curriculum and assessed the menu of offerings available in a particular classroom. Such a high number of observations are likely needed when the day-to-day instruction in reading and global teaching behaviors are highly variable, with components

receiving different levels of emphasis on any given day. With an intervention that prescribes a constant instructional model, such as the intervention being evaluated, fewer instances are likely needed to capture an estimate of a teacher's typical practice. In addition, common practice in intervention research is to observe on two to three occasions.

Methods 5: Classical test theory was applied to determine the stability of implementation fidelity ratings across occasions. An intra-class reliability coefficient was calculated based on the following equation:

$$\rho_{xx'} = \frac{\sigma^2_t}{\sigma^2_t + \frac{\sigma^2_{to,e}}{n_o}}$$

Classical test theory has several limitations. For example, only individual sources of variance and error variance are assessed and the theory assumes parallel measurement (i.e., the mean of teacher implementation would be the same across occasions). Another method, generalizabilty theory, enables researchers to assess variation from multiple sources or "facets" simultaneously, such as raters and occasions and does not assume parallel measurement.

However, because of design limitations, classical test theory was the best choice in this instance. This study is analogous to a single-faceted, nested design in generalizability theory. Specifically, occasions were nested within teachers because teachers recorded different lessons on different days. Raters could not be included as a facet because they were neither crossed with nor nested within classrooms. In a single-

facet, nested design, the G-coefficient and the reliability coefficient are analogous (Shavelson & Web, 1991). Results include significance test results from the intraclass correlation coefficient.

Research Question 6

What innovations do teachers make to the program and what are the instructional characteristics in classrooms with above average comprehension outcomes?

Hypothesis 6. Teachers' instruction will include adaptations and innovations to the model as prescribed.

Rationale 6. The criteria used to determine teachers' implementation fidelity were developed by program creators. Thus, it may be possible that some teachers who developed innovations or implemented practices not included as part of the program model were rated low on fidelity yet still provided quality instruction. Qualitative research on program implementation has demonstrated that some teachers will adapt or change the program's instructional practices. For example, Taylor and Teddlie (1999) found that teachers implementing a Title I program often omitted instructional practices outlined in the schoolwide plan. In addition, Boardman and Woodruff (2004) found that implementation fidelity is impacted by the value a school places on state assessments and teachers select components to implement or adapt based on how they perceive the practices will affect assessment results.

Methods 6. To examine instances of innovation, case studies were conducted in several stages. First, teachers who had a quality rating two-points higher than a fidelity rating were identified. Recordings from teachers who fit this criteria were examined to identify how these teachers changed or enhanced the program. Secondly, a sample of

recordings across teachers, regardless of ratings, was examined to identify patterns of innovation. Lastly, a sample of classes with post-test means above the population average were examined more closely to characterize the type of instruction provided.

Instructional characteristics were examined using case-study methodologies that included a cross-case analysis of instruction (Miles & Huberman, 1994). Specifically, patterns and themes in implementation and innovation are reported, and both common and unique innovations are described.

Chapter 4: Results

A total of 309 students and 16 teachers were included in the HLM analyses. Teacher-level data were complete, however, approximately 24% of student cases had missing data. Thus, for all HLM analyses, missing data were imputed using NORM (Schafer, 1999), a freeware program for conducting multiple imputation of incomplete multivariate data. Data were treated as missing at random because analyses conducted by researchers on the larger study revealed a lack of correlation between missingness patterns and any demographic characteristics or test scores (Simmons, Rupley, Wilson, Edmonds, & Vaughn, 2007). For this study, 10 datasets imputed in NORM were imported to HLM for analyses. The decision to impute 10 datasets was informed by guidelines reported by Shafer and Olsen (1998), the relatively low rate of missingness and the few patterns of missingness (n = 3). Imputed data were only used for HLM analyses. Other analyses, including the computation of descriptive statistics, were conducted using list-wise deletion with the original dataset.

Comprehension achievement

Descriptive statistics for the dataset used in the HLM analyses, are provided in Table 3. Student-level pre- and post-tests were significantly correlated (r = .756, p < .001), as were ratings of teacher fidelity and quality (r = .754, p < .001). Note that descriptive statistics were computed using the original dataset with missing values. Therefore, in Table 3 N represents the number of complete cases for each variable.

Table 3

Descriptive statistics from the dataset

	Variable Name	*N*	*Minimum*	*Maximum*	*Mean*	*sd*
Student-level variables						
GMRT Posttest	Y_{ij}	291	3	47	24.68	10.241
GMRT Pretest	$(Pre)_{ij}$	247	1	45	19.11	9.599
Teacher-level variables						
Fidelity rating	$(FID)_j$	16	2	7	4.54	1.551
Quality rating	$(TE)_j$	16	1	7	4.54	1.344
Class average pretest	$(PRE)_j$	16	12	29	18.97	5.109

An unconditional model with the GMRT as the outcome variable was tested to obtain summary statistics. The overall posttest average on the comprehension measure (GMRT) was 24.57 (possible range = 0 to 48). Results indicated significant variance among classrooms in comprehension achievement ($\tau = 14.14$, $p < .001$). Approximately 14% of the variance on comprehension posttest scores was between classes.

Role of fidelity and quality on student outcomes

Next, a random-coefficient model with the pre-post slope and mean achievement allowed to vary across classrooms was analyzed. In this model, the comprehension posttest average and average slope across classrooms were 24.61 and .79, respectively. Results from the random-coefficient model indicated significant variance in the model

intercept across classrooms ($\tau_{00} = 16.9$, $p < .001$), but no significant variance in slopes ($\tau_{11} = .0008$, $p > .500$). However, the full model including the implementation variables of interest (fidelity and quality) was examined before eliminating variables or modeling parameters as fixed instead of random.

Table 4 presents results for both the fixed and random effects in the fully conditional hierarchical model. Although fidelity was positively related to class mean comprehension achievement, the relationship was small and not significant. In other words, teachers with above average fidelity ratings demonstrated slightly higher, but not significant, mean comprehension achievement (controlling for quality, class-pretest and the interaction of fidelity and quality). Likewise, although instructional quality was positively related to mean classroom comprehension achievement, the relationship was close to zero and not significant. Pre-test scores aggregated at the class level did predict outcomes, indicating that classes with pretest scores above the sample average performed significantly better on the outcome.

There was a positive and significant relationship between a student's pre-test and posttest comprehension scores. However, fidelity and quality did not predict the within –school slopes. In other words, teachers with high fidelity or high quality ratings did not differ from their colleagues in the strength of relationship between pre- and post-test. Although not significant, there was a negative relationship between the slope and the implementation variables, indicating that teachers with higher fidelity and quality weakened the relationship between a student's pretest and posttest.

Table 4

Results from Fully Conditional Model

Fixed Effect	Coefficient	*se*	*t*-ratio	*df*	*p*-value
Model for class means					
Overall mean achievement (intercept), γ_{oo}	24.58	0.58	42.04	11	.000
Fidelity rating, γ_{o1}	1.25	1.44	.87	11	.404
Quality rating, γ_{o2}	.11	1.03	.11	11	.914
Class Pretest Average, γ_{o3}	.83	0.14	5.79	11	.000
Fidelity x Quality interaction, γ_{o4}	-.12	0.25	-.49	11	.637
Model for pre-post slopes					
Overall slope (intercept), γ_{10}	.81	0.06	12.92	11	.000
Fidelity rating, γ_{11}	-.17	.16	-1.08	11	.304
Quality rating, γ_{12}	-.03	.10	-.28	11	.782
Class Pretest Average, γ_{13}	-.01	.02	-.39	11	.701
Fidelity x Quality interaction, γ_{14}	.03	.03	.93	11	.372
Random Effect	Variance		χ^2		
Between Class, u_{0j}	1.71		19.57	11	.051
Between Class, u_{1j}	.001		10.34	11	> .500
Within Class, r_{ij}	44.25				

Because there was no significant variance in slope across classrooms in the original, unconditional model and none of the variables of interest were significantly related to slope in the full model, a revised model that did not allow the slope to vary across classrooms was tested. Like the full model, none of the variables of primary interest were significant (see Table 5). Likewise, students' pretests and class pretest mean continued to significantly predict comprehension outcomes.

Table 5

Results from the revised model with fixed slope

Fixed Effect	Coefficient	*se*	*t*-ratio	*df*	*p*-value
Model for class means					
Overall mean achievement, γ_{oo}	24.59	0.58	42.27	11	.000
Fidelity rating, γ_{o1}	1.25	1.44	.86	11	.407
Quality rating, γ_{o2}	.12	1.03	.12	11	.909
Class Pretest Average, γ_{o3}	.83	0.14	5.80	11	.000
Fidelity x Quality interaction, γ_{o4}	-.12	0.25	-.48	11	.638
Model for slopes					
Overall slope, γ_{10}	.79	0.05	15.36	303	.000
Random Effect	Variance		χ^2		
Between Class, u_{0j}	1.73		19.60	11	.051
Within Class, r_{ij}	44.14				

Next, the most complex term (the fidelity and quality interaction term) was deleted from the model, followed by the deletion of the least significant predictor (quality ratings). These two intermediate steps did not change results. As a final step, the remaining variable that was not significant in the revised models (fidelity ratings) was eliminated, resulting in a final model for the dataset. The final model indicates the influence of students' pretest comprehension skills on the posttest (see Table 6). In addition, there was significant variance in the outcome remaining, indicating that other variables not included in this study may explain variation in posttest scores across classrooms.

Table 6

Results from the final model

Fixed Effect	Coefficient	*se*	*t*-ratio	*df*	*p*-value
Model for class means					
Overall mean achievement, γ_{oo}	24.52	0.57	46.45	14	.000
Class Pretest Average γ_{o1}	.76	0.11	6.69	14	.000
Model for slopes					
Overall slope, γ_{10}	.79	0.04	21.30	306	.000
Random Effect	Variance		χ^2		
Between Class, u_{0j}	1.85		25.99	14	.026
Within Class, r_{ij}	44.00				

As an ancillary analysis, I conducted a simple regression to determine if teachers who taught in a departmentalized setting (i.e., they taught social studies to multiple classes) had an advantage over their colleagues in self-contained classrooms when being rated on implementation fidelity. Teacher fidelity was regressed on class-type. Class-type was not a significant predictor ($\beta = .107$, $p = .69$) indicating that departmentalized teachers did not benefit from a practice effect.

Fidelity and quality interactions

As indicated in the results reported above, no significant interaction between fidelity and quality on comprehension scores was detected. Although the relationship between the interaction of fidelity and quality and both the mean outcome and mean slope was positive, the relationship was not significant. This indicates that a teacher's level of implementation did not interact with instructional quality to result in differential effects on comprehension achievement or the pre-posttest relationship.

Correlations between teacher and researcher accounts of implementation

In addition to the HLM analyses, I examined issues related to the measurement of implementation. Of specific interest was the relationship between ratings from two different sources—teachers' logs and researcher ratings. In this part of the study, the average length of session and average fidelity score from the two sources, teacher logs and researcher coded lessons, were correlated. It is important to note that the fidelity score used in the HLM analyses was not the same score used in the correlation analyses. For HLM analyses, researcher ratings of implementation on a 7-point scale were used. These rating were informed by researchers' ratings of implementation for each

intervention component. Teachers, on the other hand, simply indicated on their logs whether a component was implemented, but did not rate implementation on any type of scale. Thus, the researcher data had to be transformed such that each component was scored dichotomously (present/absent). To obtain a fidelity score, regardless of the source, lesson components (n = 5) were dichotomously scored as either present or absent (1/0). Components were then summed to create a fidelity score for each lesson.

Session length and fidelity scores from the researcher-coded tapes were simple averages across the recordings. Average length of session and fidelity scores from teacher logs were computed from data for Weeks 5 and 6 of each Module, the two weeks during which teachers recorded the lesson that researchers coded. Fifteen teachers completed logs. Of those, some teachers failed to complete a log for every lesson. Therefore, lessons with no data were excluded when computing averages.

Fidelity scores. Both weighted fidelity scores, which placed a heavier weight on lesson components deemed critical to the intervention, and an unweighted score were correlated. The correlation between the weighted fidelity scores was moderate and significant, r = .59, p = .02. Likewise, the correlation between the unweighted fidelity scores was moderate and significant, r = .66, p = .008.

A paired samples t-test was conducted as a follow-up analysis and revealed no significant differences between the mean scores in the weighted ($t(14)$ = -1.21, p = .248) and unweighted ($t(14)$ = 1.10, p = .290) conditions. See Table 7 for the means and standard deviations for each condition.

Table 7

Means and standard deviations for fidelity scores

Condition	Researcher coded lessons (n = 15)		Teacher logs (n = 15)	
	M	SD	M	SD
Unweighted (scale of 0-5)	3.68	(1.20)	3.97	(1.04)
Weighted (scale of 0-9)	7.61	(1.32)	7.22	(1.61)

Length of session. The correlation between the average length of session reported by teachers in logs (M = 32.25, SD = 5.29) and recorded by researchers (M = 33.23, SD = 4.84) was very small and negative (r = -.08) and not significant (p = .77). However, a paired-samples t-test revealed no significant difference between the mean length of session from the two sources, $t(14)$ = -.52, p = .614.

Reliability of fidelity and quality ratings

To assess the efficiency of collecting implementation data on only a few occasions, I examined the stability of researcher ratings of fidelity (the ratings on the 7-point scaled used in HLM analyses) and quality across two measurement occasions. For this analysis, an intraclass correlation (ICC) was used. The ICC for the fidelity ratings and the overall quality ratings are reported in Table 8. The ICC for a one-way ANOVA model is an index of exact agreement, not a measure of consistency. These correlations can be interpreted as the degree of absolute agreement for independent ratings on randomly selected objects, in this case teachers (McGraw & Wong, 1996). Guidelines for

interpreting Kappa reliability coefficients (Landis & Koch, 1977) can be used to interpret ICCs. Under these guidelines, coefficients of .40 to .59 are considered moderate and .60-.79 substantial. ICCs for both a single measure and an average measure of fidelity and quality are reported. Because the average fidelity and quality rating across occasions was used in the HLM analyses reported previously, the following highlights results from the ICC for an average measure of the variable of interest. The formula used to obtain the 95% confidence interval for the average measure ICC was (SPSS, 2007):

$$\frac{F_{p/w} - F_{\alpha/2,W-1,W(k-1)}}{F_{p/w}} < \rho_{(k)} < \frac{F_{p/w} - F_{1-\alpha/2,W-1,W(k-1)}}{F_{p/w}}$$

where p is the between person effect, w the within person effect, W is number of persons, k is number of occasions, and $F_{pw} = MS_b/MS_w$.

The ICC for instructional quality ratings was substantial (ICC = .74) for an average measure across two occasions. As indicated by the confidence interval, the Average Measure ICC for quality is significantly different from zero. While this may not be particularly meaningful, as one would expect a non-zero correlation, it does provide some measure of confidence for the estimates of teacher quality obtained. The Average Measure ICC for fidelity ratings was moderate, ICC = .59. However, the confidence intervals for the fidelity rating ICCs spanned zero.

Table 8

Intraclass correlations for fidelity and quality ratings

	Intraclass Correlation	95% Confidence interval
	Fidelity rating (n = 15, k = 2)	
Single Measure	.42	-.08, .76
Average Measure	.59	-.18, .86
	Quality rating (n = 15, k = 2)	
Single Measure	.59	.15, .84
Average Measure	.74	.25, .91

n = number of teachers

k = number of occasions

Innovations to the program model

To supplement the quantitative analyses conducted as part of this study, a qualitative analysis of teachers' instructional practices was conducted. The qualitative analysis examined the extent to which teachers adhered to the program developers' vision of instruction and the types of innovations made to both procedures and strategy instruction.

Cases of low fidelity and high quality ratings. Two teachers had a quality rating that was two points higher than their fidelity rating. Both teachers were judged to provide instruction of average quality—both had a mean quality rating of 4.5 on a 7-point scale. An examination of select lessons revealed that these teachers' low fidelity scores were largely due to the absence of pre-reading activities (e.g., preteaching proper nouns,

previewing and providing the Big Idea). In place of pre-reading activities, one teacher reviewed the targeted TEKS objectives, a practice that may have been influenced by a high-stakes testing environment.

In addition, one of these two teachers had low fidelity scores on items related to teaching the target strategy. This teacher asked students to review the strategy steps on their own or simply told them to apply a strategy without modeling, explaining or reviewing *how to* apply the strategy. Another adaptation included having students read aloud round-robin style and, instead of having students practice the main idea strategy in pairs, the teacher would provide the main idea or elicit it through teacher-generated questions, calling on select students to answer.

Modifications to lessons. A sample of lessons across teachers (64% of recordings) was examined to identify modifications to the program. Across teachers, most innovations had to do with procedures rather than the comprehension strategies themselves. Some example adaptations included:

- having students write a gist for an assigned paragraph instead of each paragraph,
- substituting other types of reading (e.g., Round Robin) for partner reading,
- defining vocabulary words instead of proper nouns during pre-reading,
- reviewing content without connecting it to new text (previewing), and
- practicing strategy application as a whole class instead of in student pairs.

As the reader may recall, one component of the program was a question-generating strategy, for which there were 3 question types. A couple of teachers adapted the Question Types definitions. For example, a Level 3 Question was purposely called

"Making Connections" because we wanted students to stay with the text to answer the question. That is, we wanted them to make connections between what they just read and what they already knew. Changes to this level of question included calling it an "On your own" question that would not require any knowledge of text (e.g., How would you feel if you were an explorer?). This adapted definition is more aligned with the original Question-Answer-Relationship question types (Raphael, 1986), so teachers may have been confusing previously learned question types with the ones featured in our program.

The most prevalent adaptations were of teachers reverting to traditional ways of teaching social studies and comprehension. A typical example was characterized by having students read the text aloud (or the teacher would read the text to students), followed by teacher-generated questions, with a focus on recalling pertinent content. Such practices are more akin to what Durkin (1979) characterized as "assessing" students' knowledge rather than facilitating their ability to construct the meaning of text. In general, it appeared that low overall fidelity ratings were often a result of the *absence* of elements rather than inadequate implementation or innovations to the practices. Most frequently absent were the pre-reading activities of Previewing, Preteaching Proper Nouns, and Providing the Big Idea.

Teachers with above average outcomes. Seven teachers with class post-test averages above the overall mean were identified. Post-test averages in the identified classrooms ranged from 27.07 to 32.09. Table 9 provides information on these teachers' averages and fidelity and quality scores. Of the 7 teachers, 5 had above average fidelity scores and 6 had above average quality scores.

Table 9

Descriptive information for teachers with the highest class posttest averages

Teacher	Class Posttest Average (GMRT)	Mean Fidelity Score*	Mean Quality Score
A	27.07	5	5
B	27.43	7	6
C	27.74	5	4
D	27.93	3	5
E	28.40	7	7
F	30.50	6	6
G	32.09	3	5

* range for fidelity and quality scores = 1-7

Of most interest is Teacher G, who had the highest class average, but low fidelity scores and above average quality scores. Although this teacher did not implement many of the interventions steps, her instruction was well paced and she monitored student learning throughout the lesson. However, her strategy instruction was deemed less than adequate, and she did not hold students accountable for applying the strategies independently. For example, she would pose questions to students instead of having them generate questions or she would ask the group to tell what the text was about without individual students or pairs generating a main idea statement on their own. Her instruction was characterized by whole-class discussion of content with teacher-led questioning.

Teachers with the next highest class averages also were among the most highly rated in terms of fidelity and quality. These teachers implemented the intervention above expectations and their instruction featured many aspects of high quality strategy instruction and general features of effective instruction such as excellent lesson management and corrective feedback with guidance.

Chapter 5: Discussion

A major goal of this study was to determine whether implementation and teacher quality are related to student outcomes within a comprehension intervention program. A secondary goal was to examine different means of collecting this type of data. To address the first goal, HLM was used to model the effects of teacher quality and implementation fidelity on student outcomes. To address the secondary goals, the correlation between teacher-provided and researcher coded implementation data was examined, and the stability of fidelity and quality ratings across occasions was determined. To provide illustrations of classrooms practices quantified in the fidelity ratings, a qualitative analyses of teachers' instruction was conducted.

Using implementation data to explain outcomes

Results from this study provide several insights into the effect of implementation on student outcomes and how to approach the task of determining fidelity. First, the two indicators of implementation—fidelity and overall teacher quality—did not predict student outcomes. This differs from previous studies that have shown overall ratings of instructional quality to be related to student outcomes (e.g., Foorman et al., 2006; Taylor et al., 2002). However, it is important to note that the outcome measures used in this study, a standardized test of reading comprehension, may not have been sensitive enough to detect the effects of implementation and quality. A more proximal measure of students' knowledge and use of targeted strategies may have been better suited to modeling the effects of fidelity and quality. In addition, because taught strategies were practiced and applied while reading the district's social studies text, a measure that includes expository passages of similar content and structure may better replicate the

conditions under which students are used to applying the strategies and not require as much generalization.

In the tested model, a student's comprehension skills upon entering 4th grade did more to predict post-intervention comprehension achievement than did the teacher's instructional practices. The impact of initial comprehension skills was strong compared to the impact of teacher ratings on fidelity and quality. This appears to support an immutable view of reading achievement (Juel, 1998) —a view that places less emphasis on the role of a teacher's instruction in changing reading achievement. However, it would be errant to generalize from these results that implementation does not matter. Although we approached fidelity in a manner similar to numerous other studies—rating the presence and completeness of activities—it appears that this approach may not measure aspects of instruction that differentiate between more and less effective implementation. More sophisticated approaches to assessing fidelity that look at the nuances of a program's implementation may be needed to explain differences in student outcomes.

Given the results of previous studies that demonstrated the positive effect of overall instructional quality on student outcomes, perhaps a measure of instructional quality specific to intervention components is needed. The field has yet to identify what these nuances are or how they can best be measured, especially in a relatively complex domain of reading such as comprehension. While the traditional approach to determining fidelity used in this study may suffice for establishing internal validity, a more sophisticated approach to thinking about implementation is likely needed to model the effects on student outcomes. Although the field increasingly defines teacher quality and

effectiveness in terms of student outcomes, using instructional quality data to predict achievement remains a challenging task.

In this study, the program theory served as the foundation on which we defined fidelity. While many evaluations aim to determine program outcomes, theory-driven evaluation attempts to understand those outcomes by modeling the effects of implementation as defined by the underlying theoretical model of how the program is hypothesized to affect change. Although in the tested model measures of fidelity and quality did not explain variation in outcomes, the goal of theory-driven evaluation is valuable and worth pursuing using alternative approaches to assessing the implementation of the program's theoretical components.

Measuring fidelity and instructional quality

Despite the lack of a significant relationship between instructional quality and student outcomes in the studied model, an overall measure of teacher quality appears to be relatively reliable across only a few measurement occasions. In addition, although rating teacher quality is perceived as a relatively high inference task, the stability of the score across two occasions resembles the performance that would be expected of a low inference variable. However, there are important conditions under which this was accomplished. All of the raters involved in determining teacher quality had extensive classroom and observation experience, allowing them to make consistent judgments about instructional quality. Although not formally tested in this study, this rater characteristic appears to be important and has been a feature of all previous studies in which teacher quality was used as an explanatory variable.

Fidelity scores were less stable across occasions. Although it was hypothesized that few occasions would be needed to obtain a stable rating of fidelity given the relative simplicity of the intervention and the consistency in lesson content and routine, results suggest otherwise. Due to design limitations, the number of occasions needed to obtain a stable estimate of fidelity could not be determined. However, the literature suggests that adding even one or two more occasions may not be sufficient. These results indicate that evaluators and field-based researchers may need to rethink how best to go about determining implementation fidelity in a cost-effective manner when a large number of participants are to be observed. Determining the number of occasions required will be a function of the teacher characteristics being observed, with stable estimates of higher inference and low frequency variables requiring more observations than low-inference, frequently occurring behaviors. If more sophisticated measures of fidelity are required, as suggested in the discussion above, it is likely that researchers will need to plan for a larger number of observations as suggested by the top of the range found in the literature.

Alternative methods for collecting implementation data

The alternative method of collecting implementation data used in this study (audio recordings) appears to offer a viable and less costly means of obtaining implementation data. Recordings were clear and the amount of classroom conversation and interaction captured was beyond our initial expectations. We were able to determine implementation fidelity and overall teacher quality without the aid of "seeing" what was happening in the classroom. Data collection was efficient, and the audio files allowed us to revisit the recordings for additional analyses.

There are some concessions made when using this methodology. For one, any visual component of the intervention cannot be rated. For example, our intervention contained a graphic organizer that students used to help them write summaries. Although that component was not used to assess fidelity, had it been deemed important to the intervention, we would not have been able to determine its presence or absence. In addition, we learned that teachers not familiar with using this technology need guidelines on how best to capture the highest quality recording. For example, we had to remind teachers not to leave the recording in one place, but to carry it around with them on a lanyard and to select a lesson that was not solely independent student work. However, given the cost benefit, these concessions were deemed minor enough to continue using this methodology in future studies.

Data sources

The results from the correlations between the two sources of data—teacher logs and researcher recordings—supported what others have found (Camburn & Burns, 2004; Smithson & Porter, 1994). That is, when measured at a macro level, implementation of intervention components reported by two sources (teacher and researcher) are moderately correlated. Although this suggests that teacher logs could replace researcher-coded observations or recordings for determining *what* was taught, self-reported logs cannot be used to measure more inferential items such as overall instructional quality. Thus, depending on the needs and goals of the researcher, logs may be a viable alternative to observations or recordings. In addition, as discussed previously, measuring fidelity at a macro level may not be sufficient to model the effects of implementation nuances on student outcomes.

Although there was essentially no correlation between teacher logs and researcher codings on the length of session, there are possible explanations for this apparent anomaly. Teachers may have underreported the time spent on intervention activities to appear to comply with the program's design (a type of Hawthorne Effect if you will). Teachers were informed that lessons were designed to last approximately 30 minutes. However, through conversations with teachers and by listening to the recordings, we learned that lessons often ran longer. As reported, when length of session was aggregated across lessons, there was no difference in the estimated time between researchers and teachers, suggesting that to obtain an estimate of "average dosage" either source may be sufficient.

Implementation profiles

The implementation profiles written as part of the qualitative analyses may be excellent resources for developing hypotheses about the nuances of instruction that could be measured to better model the effects of varying implementation on student outcomes. In this sample, teachers with higher than average quality ratings were actively engaged in instruction throughout the lessons. More specifically, lesson goals and strategy explanations were clearly stated and regardless of the grouping format (whole class or student pairs), the teacher actively monitored student learning. In most of these classrooms, the teacher held students accountable for correctly applying the strategy by providing feedback with guidance and support. This contrasted with less highly rated teachers' who seldom monitored student understanding or progress during student work time or whose feedback consisted of telling students when they were wrong with little feedback as to how to correct the misapplication or misunderstanding.

Future research

Although results from this study were mixed in terms of supported hypotheses, there were lessons learned from all findings. Measuring implementation and teacher quality in a way that can be used to predict student outcomes is a challenging task for researchers and one for which there are few well-substantiated guidelines. This study may serve as a starting point from which others can develop new, innovative approaches to modeling the effects of implementation. Such approaches may be to measure the nuances that characterize the instruction of effective teachers.

Limitations

There were several limitations to this study. First, the number of Level 2 units was small compared to most applications of hierarchical modeling. As a result, the analyses were likely somewhat underpowered. Before drawing definitive conclusions about the value of the approach to modeling the effects of implementation on student outcomes examined in this study, a larger scale study may be needed. Secondly, the sample size limited the number of explanatory variables that could reasonably be entered into the model. As indicated, there was still significant remaining variance in comprehension achievement, suggesting that other factors may have explained this variation across classrooms. There may be other variables related to implementation that may be more powerful predictors of student outcomes. Thirdly, the outcome measure was a general measure of comprehension that may not have been sensitive to variations in implementation. In addition, the intervention was 18 weeks long, which may not have been long enough to detect the impact of implementation and quality on student learning. Also, teachers selected the lessons to record, introducing a variant of self-report to what is usually a random sample of instruction caught by an unannounced classroom observer.

Lastly, a single-level imputation model was used to impute datasets for the HLM analyses. While parameter estimates from single- and multi-level imputation models do not change, the standard errors in a multi-level analysis model that uses data from a single-level imputation model are somewhat biased, resulting in a higher risk of Type I errors. However, given the results of the study, it is unlikely that using a multi-level imputation model would change any study conclusions.

Appendix A

Consent Forms

IRB APPROVED ON: 08/11/2006 ***EXPIRES ON: 04/05/2007***

Examining the quality of expository text instruction and comprehension through content and case situated professional development

Conducted By: Dr. Sharon Vaughn, Ph.D. **IRB PROTOCOL # 2005040097**
University of Texas at Austin: *Vaughn Gross Center for Reading and Language Arts*; 512-232-2320; srvaughnum@aol.com

You are being asked to participate in a research study. This form provides you with information about the study. The person in charge of this research will also describe this study to you and answer all of your questions. Please read the information below and ask questions about anything you don't understand before deciding whether or not to take part. Your participation is entirely voluntary and you can refuse to participate without penalty or loss of benefits to which you are otherwise entitled. You can stop your participation at any time by simply telling the researcher.

The purpose of this study is to help develop teachers' social studies instruction and students' social studies and vocabulary knowledge and reading comprehension.

If you agree to be in this study, we will ask you to do the following things:

- Attend 3 one-half day professional development sessions (substitutes will be provided to cover teachers' instructional responsibilities).
- Allow university project personnel to audiotape social studies instruction on 3 occasions throughout the year.
- Implement 3 social studies instructional cases as part of your instruction between September and March. A case consists of teaching strategies that are presented during the professional development sessions.
- Attend 3 90-minute teacher study team sessions outside of school time between October and March.
- Complete demographic and evaluation forms rating professional development activities.

Total estimated time to participate in study is 40 hours for one academic year.

Risks and Benefits of being in the study

- There is a minimal risk of **psychological stress due to classroom observations.**
- There is a minimal risk of **loss of confidentiality**.

The potential benefits of participating in this project are:

- deepened professional knowledge in social studies instruction and
- opportunities to attend professional development that may improve teaching effectiveness.

Compensation:
You will be compensated $750.00. Compensation will occur in two installments, one at the end of each semester (December and May). If you withdraw from the study, you will be compensated for the time expended at the point of withdrawal. You will be reimbursed at state rates for mileage costs incurred for research meetings and activities.

The **records** of this study will be stored securely and kept private. Data will be entered and analyzed using ID numbers only. All measures will be securely stored in locked files in the office space of the primary investigator. Authorized persons from The University of Texas at Austin and the Institute of Education Sciences (the sponsoring agency) and members of the Institutional Review Board have the legal right to review your research records and will protect the **confidentiality** of those records to the extent permitted by law. All publications will exclude any information that will make it possible to identify you as a subject.

srvaughn

IRB APPROVED ON: 08/11/2006 ***EXPIRES ON: 04/05/2007***

Contacts and Questions:

If you have any questions about the study please ask now. If you have questions later or want additional information, call the researchers conducting the study. Their names, phone numbers, and email addresses are at the top of this page.

If you have questions about your rights as a research participant, please contact Lisa I. Leiden, Ph.D., Chair, The University of Texas at Austin Institutional Review Board for the Protection of Human Subjects, (512) 471-8871.

You will be given a copy of this information to keep for your records.

Statement of Consent:

I have read the above information and have sufficient information to make a decision about participating in this study. I consent to participate in the study.

Signature:_____ Date: _____

_____ Date: _____
Signature of Person Obtaining Consent

Signature of Investigator:_____ Date: _____

srvaughn

APPROVED BY IRB ON: 08/11/2006 EXPIRES ON: 04/05/2007
Vaughn Gross Center for Reading and Language Arts
The University of Texas at Austin · College of Education SZB 228
1 University Station D4900 · Austin, Texas 78712-0365
512-232-6565 Phone · 512-232-2322 Fax
info@texasreading.org · www.texasreading.org

Parent Copy:
Keep for your records

Enhancing the quality of expository text instruction and comprehension through content and case-situated professional development

Your son or daughter is invited to participate in a study of how professional development enhances teachers' social studies and reading instruction. My name is Sharon Vaughn, Ph.D., and I am a professor at The University of Texas at Austin, and I work with the Vaughn Gross Center for Reading and Language Arts. The purpose of this project is to learn more about how teachers' professional development improves instruction and student learning in social studies and reading comprehension. I am asking for permission to include your son or daughter in this study because his/her teacher is participating in the study and he/she has been randomly selected to be included. I expect to have up to 600 students participating this year. This research project has the support of the school principal and school officials and is funded by U.S. Department of Education, Institute of Education Science.

If you allow your child to participate, your child will be tested twice, once at the beginning and once at the end of the study, by researchers and graduate students experienced in working with students your child's age. Each testing session will take about 45 minutes to one hour.

A time will be selected so that it fits the teacher's schedule and does not require your child to miss any important information. The tests will enable us to see the effects of teaching practices on your son or daughter's comprehension, vocabulary knowledge, knowledge of social studies topics, and reading motivation. There will be no additional instruction or intervention as part of this study. All students who participate will receive the regular amounts of social studies instruction as directed by the district's curriculum and standards.

Any information obtained from this study that could identify your child will be kept strictly confidential. Information and test results will not be a part of your child's school grades. To ensure confidentiality, your child's name will be removed from all test forms and will be replaced by a number. Further, we will keep all project information in our locked offices at The University of Texas. Following completion of the project, all materials will be destroyed. Only summarized group information will be used in reports of our project, so your child's name will never be identified. Thus, any information that is obtained in connection with this study and that can be identified with your son or daughter will remain confidential and will be disclosed only with your permission. His or her responses will not be linked to his or her name or your name in any written or verbal report of this research project.

There is minimal risk associated with this study. There may be some anxiety or fatigue associated with taking the tests. To minimize these feelings, we are selecting multiple students from each class and will give the tests over two days. We will also assure

each child prior to the testing session that tests will not part of his or her grade in school. The potential benefit of participating in this study is that your child may develop improved social studies knowledge, vocabulary knowledge, and reading comprehension skills and strategies.

If you decide your child is allowed to participate, it will not cost you anything, and you will not be provided any monetary compensation for participation.

Your child's participation in this study is entirely voluntary. You are free to withdraw your consent and stop participation in this research study at any time. If you wish to stop your child's participation in this research study for any reason, you should contact Professor Sharon Vaughn at (512) 232-2320. Throughout the study, the researchers will notify you of new information that may become available and that might affect your decision to remain in the study.

Your decision to allow your son or daughter to participate will not affect your or his or her present or future relationship with The University of Texas at Austin or Del Valle Independent School District. If you have any questions about the study, please ask me. If you have any questions later, please call me at 512-232-2320. If you have any questions or concerns about your son or daughter's participation in this study, call Lisa I. Leiden, Ph.D., Chair of the University of Texas at Austin Institutional Review Board for the Protection of Human Research Participants at (512) 471-8871.

You may keep a copy of this consent form.

You are making a decision about allowing your son or daughter to participate in this study. Your signature below indicates that you have read the information provided above and have decided to allow him or her to participate in the study. If you later decide that you wish to withdraw your permission for your son or daughter to participate in the study, simply tell me. You may discontinue his or her participation at any time.

_____
Printed Name of Son or Daughter

_____ _____
Signature of Parent(s) or Legal Guardian Date

_____ _____
Signature of Investigator Date

Vaughn Gross Center for Reading and Language Arts
The University of Texas at Austin · College of Education SZB 228
1 University Station D4900 · Austin, Texas 78712-0365
512-232-2320 Phone · 512-232-2322 Fax
info@texasreading.org · www.texasreading.org

APPROVED BY IRB ON: 08/11/2006 *EXPIRES ON: 04/05/2007*

I agree to be in a study about social studies instruction. This study was explained to my mother/father and/or guardian and they said that I could be in it.

In the study, I will be given pretests to determine how well I am doing in reading and social studies and tests at the end of the study to see how much I improved. I will be asked questions about what I do when I read after the study ends. I understand that how I do in this study will be kept confidential and that tests will not be part of my school grades.

Writing my name on this page means that the page was read (by me/to me) and that I agree to be in the study. If I decide to quit the study, all I have to do is tell my parents, my teacher or the person in charge. If I at all feel distressed while participating in the study, I will contact my school counselor.

_____ _____

Child's Signature Date

_____ _____

Signature of Researcher Date

Appendix B

Sample lesson sequence

Module 1: Generating and Answering Different Levels of Questions

Week	***Lesson #***		
Week 1	Lesson 1. Teacher model phase: *Previewing (preteach proper nouns/preview text)*	Lesson 2. Review of *Previewing*	Lesson 3. Introduce *Student Study Teams*
Week 2	Lesson 4. Teacher model phase: *Level 1 Questions*	Lesson 5. Teacher assisted phase: *Level 1 Questions* (practice in student study teams)	Lesson 6. Independent practice phase: *Level 1 Questions* (practice in student study teams)
Week 3	Lesson 7. Independent practice phase continued: *Level 1 Questions* (practice in student study teams)	Lesson 8. Teacher model phase: *Level 2 Questions*	Lesson 9. Teacher assisted phase: *Level 2 Questions* (practice in student study teams)
Week 4	Lesson 10. Independent practice phase: *Level 2 Questions* (practice in student study teams)	Lesson 11. Independent practice phase continued: *Level 2 Questions* (practice in student study teams)	Lesson 12. Teacher model phase: *Level 3 Questions*
Week 5	Lesson 13. Teacher assisted phase: *Level 3 Questions* (practice in student study teams)	Lesson 14. Independent practice phase: *Level 3 Questions* (practice in student study teams)	Lesson 15. Independent practice phase continued: *Level 3 Questions* (practice in student study teams)
Week 6	Lesson 16. Practice generating all 3 Levels	Lesson 17. Practice generating all 3 Levels	Lesson 18. Class-wide Jeopardy Game

Appendix C

Sample Student Learning Log

Unfamiliar Proper Nouns	
1.	☐ Person ☐ Place ☐ Thing/Event
2.	☐ Person ☐ Place ☐ Thing/Event
3.	☐ Person ☐ Place ☐ Thing/Event
4.	☐ Person ☐ Place ☐ Thing/Event
What do I already know about this topic?	
Make a prediction: What will I learn?	
My Questions	
LEVEL ___ 1.	
Answer:	*Provide the evidence! How do you know that?*
LEVEL ____ 2.	
Answer:	*Provide the evidence! How do you know that?*
LEVEL____ 3.	
Answer:	*Provide the evidence! How do you know that?*

Appendix D

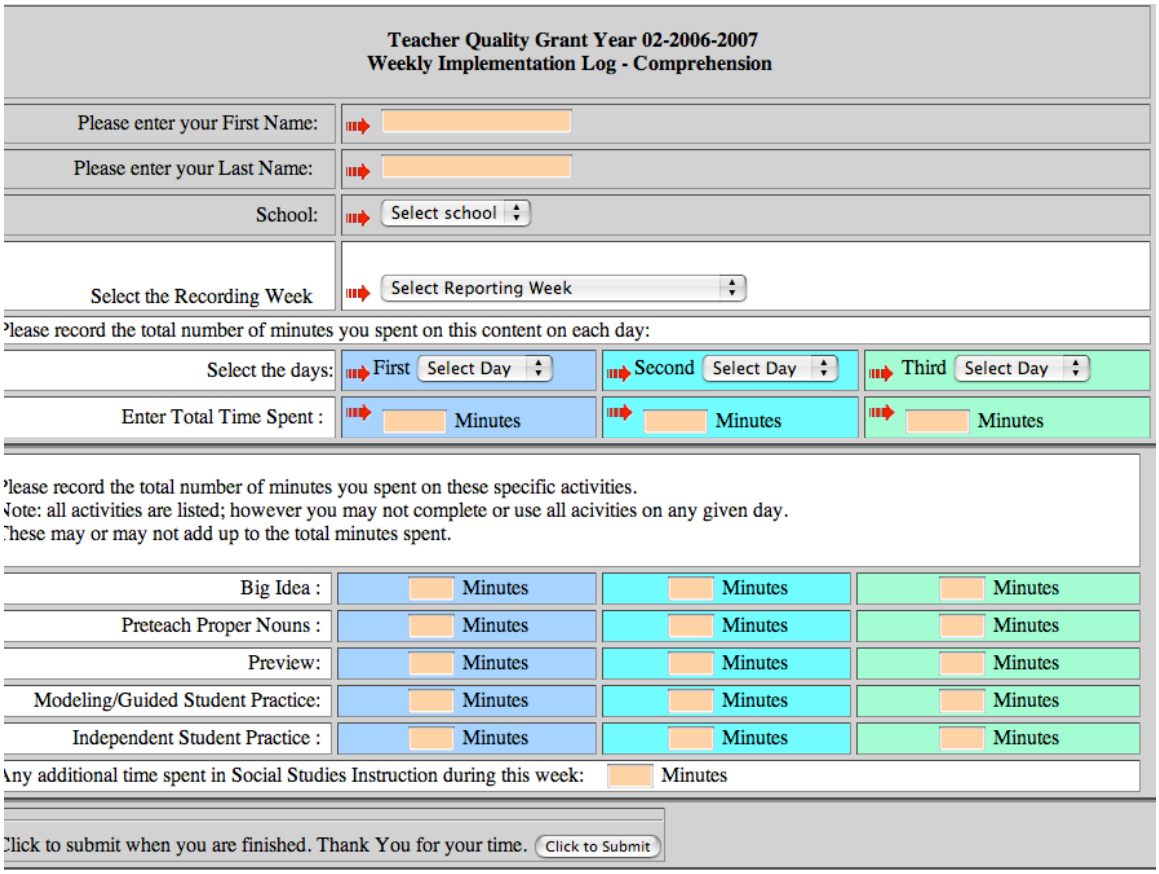
Sample Student Cue Card for Get the Gist

GET THE GIST

- ☐ What is the most important "WHO" or "WHAT" in this paragraph?
- ☐ Tell the most important idea about the "WHO" or "WHAT"
- ☐ Write your gist in ten words or less.

Appendix E

Sample Teacher Implementation Log



Teacher Quality Grant Year 02-2006-2007
Weekly Implementation Log - Comprehension

Please enter your First Name:			
Please enter your Last Name:			
School:	Select school		
Select the Recording Week	Select Reporting Week		
Please record the total number of minutes you spent on this content on each day:			
Select the days:	First Select Day	Second Select Day	Third Select Day
Enter Total Time Spent :	Minutes	Minutes	Minutes

Please record the total number of minutes you spent on these specific activities.
Note: all activities are listed; however you may not complete or use all acivities on any given day.
These may or may not add up to the total minutes spent.

Big Idea :	Minutes	Minutes	Minutes
Preteach Proper Nouns :	Minutes	Minutes	Minutes
Preview:	Minutes	Minutes	Minutes
Modeling/Guided Student Practice:	Minutes	Minutes	Minutes
Independent Student Practice :	Minutes	Minutes	Minutes

Any additional time spent in Social Studies Instruction during this week: Minutes

Click to submit when you are finished. Thank You for your time. Click to Submit

Appendix F

Comprehension Fidelity Coding Form and Codebook

Teacher Quality: Fidelity Check

Teacher name:	Round 1 2 3	Observer name:
School:	Topic:	Total Time:
Notes: (Please note any unusual or special circumstances that would affect the fidelity check.)		

Instructions:
Items 1-6: Rate each item using the 4-point scale. If you cannot determine the presence/absence of an item, rate the item as not observable (e.g., the tape was muffled).

Item 7: Use your ratings of items 1-6 to determine a rating of overall implementation.
Item 8: Consider the teaching quality overall, regardless of fidelity to the model.

COMPREHENSION

Component	Implementation				
The teacher:	**3**	**2**	**1**	**Not at all (0)**	**Not observable**
1. Preteaches/reviews proper nouns					
2. Provides the "big idea"/main idea for the selected reading					
3. Guides students through the preview					
4. Models/explains/ reviews **target** strategy					
5. Models/explains/ reviews **another** strategy					
6. Has students apply strategies					
KEY: 1 = minimal/unacceptable implementation; **2** = acceptable implementation; **3** = exemplary implementation					

Target Strategies		
Module 1	*Module 2*	*Module 3*
Asking/ answering different types/levels of questions	Writing a gist/main idea statement	Using the graphic organizer OR Writing summaries/ Applying the summary rules

Overall Ratings: Use the above ratings to inform your ratings of item 7

INSTRUCTIONAL QUALITY: Use the above ratings to inform your ratings of item 7

Overall Ratings of Teacher Implementation	**Less than adequately** ◄——————————► **Above expectations**
7. Overall, this teacher implemented the instructional practices:	1 2 3 4 5 6 7

Overall Ratings of Teacher Quality	**Not at all effective** ◄——————————► **Highly effective**
8. Overall, I consider this teacher's instruction to be:	1 2 3 4 5 6 7

Comprehension Codebook

Item	Indicator of Exemplary Implementation
1. Pre-teaches proper nouns	• 3-5 key words are pre-taught or reviewed • Both how to say them and what they mean • NOTE: Lengthy explanations of words does not equal exemplary implementation; should last only a few minutes
2. Provides the "big idea"/main idea for the selected reading	• Teacher states the Big Idea at the beginning of every lesson. • Big idea = the most important thing the teacher wants students to understand and remember from the reading
3. Guides students through the preview	• Students state what they already know about the topic • Students make predictions about what they will read • Students make connections between what they know and will read • NOTE: the entire preview including preteaching PN should take no longer than 10 minutes
4. Models/explains reviews the target strategy: (indicators apply to #5 as well)	
Module 1: Asking/ answering different types/levels of questions	• Instruction/practice focuses on students generating one or more of the types of questions (Levels 1, 2, or 3)
Module 2: Writing a gist/main idea statement	• Instruction/practice on the Gist procedure (most important who/what and most important thing about that who/what in 10 words or less)
Module 3: Using the graphic organizer OR	• Students taught to use or practice using the content web • Gists are written on the web • Big idea provided by the teacher; represented in the center oval
Module 3: Writing summaries/ Applying the summary rules	• Instruction or practice on writing summaries using the summary rules • Uses the content web to assist with summary writing
6. Has students apply strategies	• Verbal evidence that students are working in pairs or with the class (during guided practice) to apply the strategies

Item 4-5

Depending on the lesson recorded, items 4-5 may involve modeling the strategy, explaining/reviewing the strategy and providing opportunities for students to practice applying the strategy. Below are indicators of high quality practices for each stage.

Stage	Indicators
Modeling	• The teacher makes her thinking visible • Involves showing students *how to* as opposed to explaining or stating what the strategy is
Explains/reviews	• Before or after student practice, the teacher reviews the strategy, highlighting key steps • Explanation is accurate • Teacher solicits input from students to ensure understanding
Student Practice	• May be guided or paired • Students are given ample time to apply the strategies with feedback (especially corrective feedback)

Item 7:

Overall, this teacher implemented the instructional practices:

Consider the recorded lesson in its entirety, using the ratings on items 1-6 to substantiate your rating.

Item 8:

Overall, I consider this teacher's instruction:

Regardless of adherence to the comprehension/vocabulary model, rate the effectiveness of this teacher's instruction.

References

Anderson, L.M., Everston C.M., & Brophy, J.E. (1979). An experimental study of effective teaching in first-grade reading groups. *The Elementary School Journal, 79,* 190-223.

Beck, I. L., & McKeown, M. G. (2001). Inviting students into the pursuit of meaning. *Educational Psychology Review, 13*(3), 225-241.

Beck, I. L., McKeown, M. G., Hamilton, R. L., & Kucan, L. (1997). *Questioning the Author: An approach for enhancing student engagement with text*. Newark, DE: International Reading Association.

Beck, I. L., McKeown, M. G., Worthy, J., Sandora, C. A., & Kucan, L. (1996). Questioning the Author: A year-long classroom implementation to engage students with text. *Elementary School Journal*, *96*, 385-414.

Boardman, A.G. & Woodruff, A.L. (2004). Teacher change and 'high-stakes' assessment: What happens to professional development? *Teaching and Teacher Education, 20*(6), 545-557.

Borich, G. (2003). *Observation skills for effective teaching. 4th edition.* Upper Saddle River N.J.: Merrill.

Brophy, J.E. (1979). Teacher behavior and its effects. *Journal of Educational Psychology, 71*(6), 733-750.

Camburn, E. & Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal, 105*, 49-73.

Chen, H. (1990). *Theory driven evaluation*. Newbury Park, CA: Sage.

Chall, J. S. (1983). *Stages of reading development*. New York: McGraw-Hill.

Confrey, J. (2006). Comparing and contrasting the National Research Council Report *On Evaluation Curricular Effectiveness* with the What Works Clearinghouse approach. *Educational Evaluation and Policy Analysis, 28*(3), 195-213.

Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension: Examining child x instruction interactions. *Journal of Educational Psychology, 96*(4), 682-698.

Deshler, D. D., & Hock, M. F. (2007). Adolescent literacy: Where we are, where we need to go. In M. Pressley, A. Billman, K. Perry, K. Reffitt, & J. Moorhead Reynolds (Eds.), *Shaping literacy achievement: Research we have, research we need.* New York: Guilford.

Dole, J. A., Brown, K.J., and Trathen, W. (1996). The effects of strategy instruction on the comprehension performance of at-risk students. *Reading Research Quarterly, 31*(1), 62-88.

Duffy G.G. & Roehler, L.R. (1982). Matching direct instruction to reading outcomes. *Language Arts, 59,* 476-480.

Duke, N. K., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (pp. 205-242). Newark, DE: International Reading Association.

Durkin, D. (1979). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly, 14,* 481-533.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C. K., Cable, A., Tackett K., & Wick, J. (in press). A synthesis of reading interventions and effects on reading

outcomes for older struggling readers. *Review of Educational Research.*

Erlich, O. & Shavelson, R.J. (1978). The search for correlations between measures of teacher behavior and student achievement: Measurement problem, conceptualization problem or both? *Journal of Educational Measurement, 15*, 77-89.

Fletcher , J. M., Francis, D. J., Boudousqui, A., Copeland, K., Young, V., Kalinowski, S., et al. (2006). Effects of accommodations on high-stakes testing for students with reading disabilities. *Exceptional Children*, *72*(2), 136-150.

Foorman, B.R., Chen, D.T., Carlson, C., Moats, L., Francis, D.J., Fletcher, J.M. (2003). The necessity of the alphabetic principle to phonemic awareness instruction. *Reading and Writing, 16,* 289-324.

Foorman, B.R., Francis, D.J., Fletcher, J.M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*, 37-55.

Foorman, B. & Schatschneider, C. (2003).Measurement of teaching practices during reading/language arts instruction and its relationship to student achievement. In S. R. Vaughn & K. L. Briggs (Eds.) *Reading in the classroom: Systems for observing teaching and learning (*1-30). Baltimore: Paul H. Brookes.

Foorman, B.R., Schatschneider, C., Eakin, M., Fletcher, J.M., Moats, L.C., and Francis, D.J. (2006). The impact of instructional practices in grades 1-2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology, 31,* 1-29.

Frechtling, J.A., Zhang, X., and Silverstein, G. (2006). The Voyager Universal Literacy system: Results from a study of kindergarten students in inner-city schools.

Journal of Education for Students Placed at Risk, 11, 75-95.

Fuchs, D., Fuchs, L.S., Mathes, P.G., & Simmons, D.C. (1997). Peer assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34*, 174-206.

Fuchs, L.S., Fuchs, D., and Karns, K. (2001). Enhancing kindergarteners' mathematical development: Effects of peer-assisted learning strategies. *Elementary School Journal, 101,* 495-510.

Fulk, B.M. & King, K. (2001). Classwide peer tutoring at work. *Teaching Exceptional Children,34,* 49-53.

Gersten, R., Baker, S. & Lloyd, J.W. (2000). Designing high-quality research in special education: Group experimental design. *Journal of Special Education, 34*(1), 2-18.

Gersten, R., Coyne, M., Compton, D., Fuchs, L.S., Greenwood, C., & Innocenti, M. (2005). Quality Indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research, 71*(2), 279–320.

Graves, A. W. (1986). Effects of direct instruction and metacomprehension training on finding main ideas. *Learning Disabilities Research, 1*(2), 90-100.

Graves, M. F., Cooke, C. L., & Laberge, M. J. (1983). Effects of previewing difficult short stories on low-ability junior high school students' comprehension, recall, and attitudes. *Reading Research Quarterly, 18,* 262-276.

Greenwood, C.R. & Delquardi, J. (1995). Classwide peer tutoring and the prevention of

school failure. *Preventing School Failure, 39,* 21-25.

Good, T.L., & Brophy, J. E. (2000). *Looking in classrooms* (8th ed.). New York: Longman.

Hartmann, D. P. & Wood D. D. (1990). Observational methods. In Alan S. Bellack & Michel Hersen (Eds.) *International handbook of behavior modification and therapy* (2nd ed., pp. 107-138). New York: Plenum.

Henk, W.A., Moore, J.C., Marinak, B.A. & Tomasetti, (2000). A reading lesson observation framework for teachers, principals, and literacy supervisors. *The Reading Teacher, 53*(5), 358-369.

Jackson, J.B., Paratore, J.R., Chard, D.J., and Garnick, S. (1999). An early intervention supporting the literacy of children experiencing substantial difficulty. *Learning Disabilities Research and Practice, 14,* 254-267.

Jetton, T.L. & Alexander, P.A. (2004). Domains, teaching and literacy. In T.L. Jetton & J.A. Dole (Eds.), *Adolescent Literacy Research and Practice* (pp. 15-39). New York: Guilford.

Jenkins, J. R., Heliotis, J. D., Stein, M. L., & Haynes, M. C. (1987). Improving reading comprehension by using paragraph restatements. *Exceptional Children, 54*(1), 54-59.

Jitendra, A. K., Cole, C. L., Hoppes, M. K., & Wilson, B. (1998). Effects of a direct instruction main idea summarization program and self-monitoring on reading comprehension of middle school students with learning disabilities. *Reading and Writing Quarterly, 14*(4), 379-396.

Jitendra, A. K., Hoppes, M. K., Xin, Y.P. (2000). Enhancing main idea comprehension for students with learning problems: The role of a summarization strategy and self-monitoring instruction. *Journal of Special Education, 34*, 127-139.

Juel, C. (1998). *Teaching Reading in the 21st Century.* Boston: Allyn & Bacon.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis, 21*(4), 345-363.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. New York: Cambridge University Press.

Kintsch, W. & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71–92). Mahwah, NJ: Lawrence Erlbaum.

Kim, A., Vaughn, S., Wanzek, J., & Wei, S. (2004). Graphic organizers and their effects on the reading comprehension of students with learning disabilities. *Journal of Learning Disabilities, 37*, 105–118.

Kovaleski, J.F. (1999). High versus low implementation of instructional support teams: A case for maintaining program fidelity. *Remedial and Special Education, 20*(3), 170-183.

Landis, J.R. & Koch G.G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33,* 363-374.

Lysynchuk, L.M., Pressley, M., d'Ailly, H., Smith, M. & Cake, H. (1989). A methodological analysis of experimental studies of comprehension strategy instruction. *Reading Research Quarterly,24,* 458-470

MacGinitie, MacGinitie & Dreyer (2000). Gates McGinnitie Reading Test, 4th edition. Riverside Publishing.

Magano, N.G. (1982) The development and psychometric testing of a systematic observation instrument to study teacher effectiveness in the reading classroom. Unpublished dissertation; Texas A& M University.

Malone, L. D., & Mastropieri, M. A. (1991). Reading comprehension instruction: Summarization and self-monitoring training for students with learning disabilities. *Exceptional Children, 58*(3), 270-279.

Mastropieri, M. A., Scruggs, T. E., Bakken, J. P. & Whedon, C. (1996). Reading comprehension: A synthesis of research in learning disabilities. *Advances in Learning and Behavioral Disabilities, 10B,* 201-227.

Mathematica Policy Research (2006). National Study of the Effectiveness of Reading Comprehension Interventions. U.S. Department of Education, Institute of Education Sciences.

Mathes, P.G., Torgesen, J.K., and Allor, J.H. (2001). The effects of peer-assisted literacy strategies for first-grade readers with and without additional computer-assisted instruction in phonological awareness. *American Educational Research Journal, 38,* 371-410.

McCabe, P. (1992). The multidimensional reading instruction observation scale. *Reading Horizons, 33*(2), 167-176.

Miles. M.B. & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.

Mowbray C.T. Holter, M.C., Teague, G.B., Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315-340.

National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read* (NIH Pub. No. 00-4754). Washington, DC: U.S. Department of Health and Human Services.

National Research Council (2004). *On evaluation curricular effectiveness: Judging the quality of K-12 mathematics research.* Washington, DC: The National Academy Press.

Pearson, P.D. & Dole, J.A. (1987). Explicit comprehension instruction: A review of the research and a new conceptualization of instruction. *Elementary School Journal, 88,* 153-167.

Porter, A. (2002). Measuring the content of instruction: Research and practice. *Educational Researcher, 31*(7), 3-14.

Porter, A.C., & Brophy, J. (1988). Synthesis of research on good teaching: Insights from the work of the Institute for Research on Teaching. *Educational Leadership,* 74-85.

Pressely, M. (1998). Comprehension strategies instruction. In J. Osborn & F. Lehr (Eds.), *Literacy for All: Issues in Teaching and Learning*, pp. 113-133. New York: Guildford.

Pressley, M., Rankin, J., Yokoi, L. (1996). A survey of instructional practices of primary grade teachers nominated as effective in promoting literacy. *The Elementary School Journal, 96*(4), 363-384.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. Kamil, P. Mosenthal, P. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 545–562). Mahwah, NJ: Erlbaum.

Pressley, M. (2001). Comprehension strategy instruction: A turn-of the century status report. In C. Block & M. Pressley (Eds.) *Comprehension instruction: Research-based best practices*. (pp. 11-27), New York: Guilford.

Pressley, M. (2002). *Reading instruction that works: The case for balanced teaching* (2nd ed). New York: Guilford.

Raphael, T. (1986). Teaching question and answer relationships, revisited. *Reading Teacher, 39*, 516-522.

Raphael, T. E., Highfield, K., & Au, K. H. (2006). QAR now: A powerful and practical framework that develops comprehension and higher-level thinking in all students. New York: Scholastic.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models: Application and data analysis methods* (2nd edition). Thousand Oaks, CA: Sage.

Raudenbush, S.W., Bryk, A.S. & Congdon (2006). HLM Version 6 software. Scientific Studies International.

Raudenbush, S.W. & Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401.

Roe, M.F. and Vukelich, C. (2001). Understanding the gap between AmericaReads program and the tutoring session: The nesting of challenges. *Journal of Research in Childhood Education, 16,* 39-52.

Rogers, P.J. (2000). Program theory: Not whether programs work, but how they work. In D.L. Stufflebeam, G.F. Madaus, & T. Kellaghan (Eds.), *Evaluation Models: Viewpoints on Educational and Human Services Evaluation,* pp. 209-232. Boston: Kluwer Academic Publishers.

Rosenshine, B. (1980). Content, time, and direct instruction. In P. Peterson & H. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications*. Berkeley, CA: McCutchan.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of intervention studies. *Review of Educational Research, 66*(2), 181-221.

Rosenshine, B. & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64*(4), 479-530.

Rossi, P.H., Freeman, H.E., & Lipsey, M.W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.

Rowan, B. (2005). Measuring classroom practice. Presentation at the NCEE Workshop on Measuring Classroom Practice, October, 2005. Washington D.C.

Rowan, B., Camburn, E. & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third grade classrooms. *The Elementary School Journal, 105*(1), 75-101.

Saxe, G.B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction, 17*(1), 1-24.

Schafer, J.L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer, J.L. & Olsen, (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545-571.

Shavelson, R.J. & Dempsey, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46,* 553-611.

Simmons, D., Rupley, W., Wilson, V., Edmonds, M., & Vaughn, S.R. (2007). Unpublished report to the Institute of Education Sciences for grant contract number R305M050121A: *Enhancing the quality of expository text instruction and comprehension through content and case-situated professional development*.

Smithson, J., & Porter, A. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum—the Reform Up Close Study* (Research Report No. 31). Madison, WI: University of Wisconsin, Consortium for Policy Research in Education.

Snijders T.A. & Bosker, R. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks: Sage.

Snow, C. (2002). Reading for understanding: Toward an R&D program in reading comprehension. Santa Monica, CA: RAND.

Social Studies Texas, Grade 4. (2003). Glenview, IL: Scott Foresman/Pearson Education, Inc.

SPSS (2007). *SPSS 14.0 Statistical Algorithms: Reliability*. Downloaded from

http://support.spss.com/Student/Documentation/Algorithms/index.html on October 29, 2007.

Swanson, H. L. (1999). Reading research for students with LD: A meta-analysis of intervention outcomes. *Journal of Learning Disabilities, 32*, 504–532.

Taylor B., Pearson, D., Clark, K., & Walpole, S. (2002). Looking inside classrooms: Reflecting on *the how* as well as *the what* in effective reading instruction. *The Reading Teacher*, 56(3), 270-279.

Taylor, B.M., Pearson, P.D., Peterson, D.S., & Rodriguez, M.C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal, 104,* 3-28.

Taylor, D.L. & Teddlie, C. (1999). Implementation fidelity in Title I schoolwide programs. *Journal of Education for Students Placed At Risk, 4*(3), 299-319.

Troia, G.A. (1999). Phonological awareness intervention research: A critical review of the experimental methodology. *Reading Research Quarterly, 34,* 28-52.

Vaughn, S. (2002). Unpublished fidelity measure for the Three Tier Reading Intervention as a Means of Preventing Reading Difficulties. Office of Special Education Programs.

Vaughn, S. (2001). Unpublished fidelity measure for the Effect of Duration of Intensive Instruction on the Reading Progress of Struggling Readers. Office of Special Education Programs.

Weiss, C.H. (1997a). Theory-based evaluation: past, present, and future. In D.J., Rog & D. Fournier (Eds.), *Progress and future directions in evaluation: Prespective on*

theory, practice and methods: New Directions for evaluations (Issue 76, pp. 40-55). San Francisco: Jossey-Bass.

Weiss, C.H. (1997). *Evaluation*. (2nd edition). Upper Saddle River, NJ: Prentice Hall.

What Works Clearninghouse (2005). What works clearninghouse review process. Retrieved November 6, 2005 from http://www.whatworks.ed.gov/.

Wong, B. Y. L., & Jones, W. (1982). Increasing metacomprehension in learning disabled and normally achieving students through self-questioning training. *Learning Disability Quarterly, 5,* 228-240.

Yin, R.K. (1994). Evaluation: A singular craft. In C.S. Reichardt & S.F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives.* New Directions for Program Evaluation, no. 61. San Francisco: Josey-Bass.

VITA

Meaghan Suzanne Edmonds was born in Indianapolis, Indiana on October 18, 1972, the daughter of Michael and Julia Weis. After completing her work at Immaculata High School in Somerville, New Jersey, in 1991, she entered the University of Notre Dame in South Bend, Indiana. She received the degree of Bachelor of Arts from the University of Notre Dame in May 1995. During the following years she was employed as an American History teacher at Bishop Lynch High School in Dallas, Texas. In June 1998 she entered The University of Texas at Austin. She received the degree of Master of Arts in August 1999 and the degree of Master of Education in December 2005. She has been employed as a Research Associate at The Vaughn Gross Center for Reading and Language Arts at The University of Texas at Austin since September 1999.

Permanent Address: 7703 Long Point Drive
Austin, Texas 78731

This dissertation was typed by the author.